TITULNÍ LIST PERIODICKÉ ZPRÁVY 2009 PROJEKTU 2C06009

Ministerstvo školství, mládeže a tělovýchovy

2C06009

PROSTŘEDKY TVORBY KOMPLEXNÍ BÁZE ZNALOSTÍ PRO KOMUNIKACI SE SÉMANTICKÝM WEBEM V PŘIROZENÉM JAZYCE

řešitel - **doc. Ing. Karel Ježek, CSc.**

.....

(podpis)

za příjemce - koordinátor - **Západočeská univerzita v Plzni** (IČ: 49777513)

rektor

Doc. Ing. Josef Průša, CSc.

.....

(podpis, razítko)

Verze zprávy: **1** Zpracováno dne:

2. SKUTEČNOST ZA UPLYNULÉ OBDOBÍ - 2009

2.1. PROJEKTOVÝ TÝM A ŘEŠITELSKÉ TÝMY

2.1.1. PROJEKTOVÝ TÝM

IČ organizace	49777513
Obchodní jméno - název	**Západočeská univerzita v Plzni**
Zkratka názvu	ZČU
Role organizace	příjemce - koordinátor
Vazba na organizaci	00216224
Druh organizace	Veřejná nebo státní vysoká škola (zákon č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (o vysokých školách)
Adresa sídla, spojení na organizaci	
- ulice, čp./č.or.	Univerzitní 2732/ 8
- PSČ, obec	30614 Plzeň
- stát	Česká republika
- telefon	377 631 111
- http://	www.zcu.cz
Bankovní spojení	
-DIČ	CZ49777513
- banka kód, název	0100 - Komerční banka, a.s., Plzeň
- číslo účtu, sp.symbol	4811530257,
Statutární zástupce	
- titul před, jméno, příjmení, titul za	Doc. Ing. Josef Průša CSc.
- funkce	rektor
- telefon	377631000
- mobil	606665105
- fax	377631002
- email	rektor@rek.zcu.cz

IČ organizace	00216224
Obchodní jméno - název	**Masarykova univerzita**
Zkratka názvu	MU
Role organizace	spolupříjemce
Vazba na organizaci	49777513
Druh organizace	Veřejná nebo státní vysoká škola (zákon č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (o vysokých školách)
Adresa sídla, spojení na organizaci	
- ulice, čp./č.or.	Žerotínovo náměstí 617/ 9
- PSČ, obec	60177 Brno
- stát	Česká republika
- telefon	549 491 1111
- http://	www.muni.cz
Bankovní spojení	
-DIČ	CZ00216224
- banka kód, název	0100 - Komerční banka Brno-město
- číslo účtu, sp.symbol	85636621,
Statutární zástupce	
- titul před, jméno, příjmení, titul za	Prof. PhDr Petr Fiala PhD
- funkce	rektor
- telefon	549491001
- mobil	
- fax	
- email	rektor@muni.cz

2.1.2. ŘEŠITELSKÝ TÝM

Celé jméno, RČ	**Bártek Luděk Mgr.** 7201083791 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 3215 bar@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita Fakulta informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	30

Celé jméno, RČ	**Brada Přemysl Ing. PhD. MSc.** 7007012111 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	3772435 brada@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

Celé jméno, RČ	**Češka Zdeněk Ing.** 8207311244 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632452 zceska@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	100

Celé jméno, RČ	**Dostal Martin ing.** 8409092054 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632452 604796109 madostal@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	25

Celé jméno, RČ	**Ekštein Kamil Ing. PhD.** 7705302011 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 kekstein@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Fiala Dalibor Ing PhD.** 8003235845 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632429 dalfia@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	25

Celé jméno, RČ	**Habernal Ivan Ing.** 8307051764 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 377 632 402 habernal@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	50

Celé jméno, RČ	**Hejtmánek Jan Ing.** 8211012095 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 377 632 402 hejtman2@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	12.5

Celé jméno, RČ	**Horák Aleš RNDr. Ph.D.** 7409014250 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 4377 hales@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita Fakulta informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	50

Celé jméno, RČ	**Hynek Jiří ing. PhD.** 7205062029 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632455 hynekj@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Ježek Karel doc. Ing. CSc.** 420617110 CZ
Role osoby při řešení projektu	řešitel
Spojení	377 632 475 jezek_ka@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Klečková Jana doc. Dr. Ing.** 496108095 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 421 kleckova@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

Celé jméno, RČ	**Konopík Miloslav Ing. PhD.** 8103261782 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 konopik@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	100

Celé jméno, RČ	**Kopeček Ivan doc. RNDr. CSc.** 490303075 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 3861 kopecek@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita Fakulta informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	40

Celé jméno, RČ	**Král Pavel Ing. PhD.** 7603172049 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632454 pkral@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Krčmář Lubomír Ing.** 8408221228 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632452 lkrcmar@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	25

Celé jméno, RČ	**Krutišová Jana Ing.** 5955160046 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 413 krutisova@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

Celé jméno, RČ	**Matoušek Václav prof. Ing. CSc.** 480613108 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 471 matousek@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Mautner Pavel Ing. PhD.** 6505222592 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 441 mautner@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Mouček Roman Ing. PhD.** 7607072000 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 441 moucek@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Pala Karel doc. PhDr. CSc.** 390615416 CZ
Role osoby při řešení projektu	spoluřešitel
Spojení	549 49 5616 pala@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita Fakulta informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Pavelka Tomáš Ing. PhD.** 7909182083 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 tpavelka@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	100

Celé jméno, RČ	**Pomikálek Jan Mgr.** 7910090419 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 1864 xpomikal@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita Fakulta informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	10

Celé jméno, RČ	**Ptáčková Helena** 7059142079 CZ
Role osoby při řešení projektu	osoba autorizovaná k finančním záležitostem
Spojení	377 632 463 377 632 402 ptackova@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	5

Celé jméno, RČ	**Rambousek Adam Bc.** 8110225233 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	xrambous@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita Fakulta informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	50

Celé jméno, RČ	**Rohlík Ondřej Ing. PhD.** 7510031925 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632450 rohlik@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	50

Celé jméno, RČ	**Rychlý Pavel Mgr. Ph.D.** 7301235359 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 6399 pary@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita Fakulta informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	50

Celé jméno, RČ	**Sojka Petr doc. RNDr. Ph.D.** 6309171000 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549496966 sojka@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita Fakulta informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

Celé jméno, RČ	**Steinberger Josef Ing. PhD.** 7909182127 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 479 jstein@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Toman Michal Ing.** 8007042054 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632479 mtoman@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	100

Celé jméno, RČ	**Zíma Martin Ing. PhD.** 7405042073 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632431 zima@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

2.1.3. ZMĚNY V PROJEKTOVÉM A ŘEŠITELSKÝCH TÝMECH - rok 2009

Pč.	Typ	Popis
1	změny v projektovém týmu a řešitelských týmech	K 31.1.2009 odešel z týmu ing. Ondřej Rohlík, PhD. z důvodu nástupu na jiné pracoviště.
2	změny v projektovém týmu a řešitelských týmech	K 30.6.2009 odešel z týmu ing. Jiří Hynek, PhD., z důvodu práce ve vlastní firmě.
3	změny v projektovém týmu a řešitelských týmech	V srpnu 2009 ukončili práci na projektu Ing. Tomáš Pavelka, PhD. a Ing. Mochal Toman z důvodu odchodu na jiné pracoviště.
4	změny v projektovém týmu a řešitelských týmech	K 15.září 2009 nastoupili do týmu Ing. Martin Dostál a Ing. Lubomír Krčmář.
5	změny v projektovém týmu a řešitelských týmech	K 1. listopadu se do týmu vrátil ze zahraniční praxe Ing.Dalibor Fiala, PhD.
6	změny v projektovém týmu a řešitelských týmech	K 1. listopadu odešel na zahraniční výzkumné pracoviště ing. Josef Steinberger, PhD.
7	změny v projektovém týmu a řešitelských týmech	K 31.12.2009 odchází do komerční sféry ing. Zdeněk Češka, PhD.

2.2. ČASOVÝ POSTUP PRACÍ

Komentář k metodice a časovému postupu prací a průběhu aktivit za uplynulé období

Časový rozvrh, způsob a postup prací byly dodrženy. Během roku se však vyskytla potřeba doplnění několika dalších aktivit, které vyplynuly z uskutečňovaných prací a dosažených výsledků. Vzhledem k jejich rozsahu bylo vhodné je zařadit jako samostatné položky do předkládané zprávy. Jinak se plán řešitelských prací nezměnil a byl plně dodržen.

2.2.0. PŘEHLED DÍLČÍCH CÍLŮ SCHVÁLENÉ- SKUTEČNOST 2009

	Číslo	Dílčí cíl podrobně	Datum plnění
	1	**Dílčí cíl** Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřování algoritmů komunikace s www prostředím. **Indikátory dosažení - výsledky dílčího cíle** a) Vytvoření uživatelského rozhraní pro hlasový vstup / příp. výstup, které bude použito pro komunikaci se sémantickým webem, a pro jeho podporu vytvoření robustního ASR systému pro inflexní jazyky. K tomu bude nutno vytvořit kvalitní korpus pro ASR a z něj extrahovat dostatečné množství trénovacích dat. V jednotlivých etapách bude v průběhu let 2006 – 2007 vytvořen: - kvalitní audio-korpus pro natrénování systému ASR, - korpus pro natrénování jazykových modelů. b) Příprava datových kolekcí a pomocných rutin vyhledávacího systému ve vícejazyčných korpusech, včetně prostředků pro zpřesňování uživatelských dotazů pomocí thesauru a nástrojů pro disambiguaci víceznačných slov, na bázi klient/server aplikace. Jednotlivé dílčí výsledky řešení projektu lze charakterizovat takto: - vytvoření multijazykových korpusů – základní výběr zahrnuje angličtinu a češtinu, dle možností alespoň některé úlohy plánujeme provádět i se slovenštinou (zajímavá je blízkostí k češtině) a němčinou, - metoda automatického rozpoznání jazyka – kombinace „stop slov“ a frekvenčních znakových metod. c) Příprava datových kolekcí a modulů pro filtraci a sumarizaci textů: - vytvoření sumarizačních korpusů (pro angličtinu plánujeme využít standardních korpusů, např. DUC a pro češtinu bude vytvořen vlastní, složený vesměs z textů novinových článků, - sumarizace textů založená na latentní sémantické analýze (LSA), vytvoření anotované kolekce pro sumarizátor založený na LSA - vytvoření vícejazyčných korpusů, - rozšíření standardních textových korpusů o korpusy závadných dokumentů pokrývající problematická témata definovaná v zadání. d) Korpus syntaktických stromů (treebank): - korpus bude morfologicky označkován a zjednoznačněn, - bude v něm vyznačena závislostní struktura věty i jednotlivé větné složky včetně koreferenčními vztahy, - korpus bude z části založen na existujícím PDT. e) Korpus vzorových přepisů vybraných vět a jejich sémantické reprezentace: - text korpusu bude podmnožinou korpusu syntaktických stromů, - ve stromech budou vyznačeny významy z dostupných ontologií (WordNet), - věty budou rozšířeny o logické formy. f) Doplnění morfologického značkovače o robustní hádací proceduru, která bude spolehlivě přiřazovat morfologické značky i neznámým slovům. **Prostředky ověření - Forma zpracování a předání výsledku dílčího cíle** Jedná se o vytvoření podpůrného aparátu, bez něhož nelze další zamýšlené cíle projektu dosáhnout. Vytvořeny budou proto korpusy v podobě rozsáhlých datových souborů se specifickou strukturou a organizací a pro jejich údržbu a prohledávání budou vyvinuty speciální softwarové nástroje. Výsledky budou soustředěny do soustavy datových souborů a její obsah prezentován formou publikace na konferencích a v průběžných výzkumných zprávách. **Kritické poedpoklady dosažení dílčího cíle** Rizikové faktory ovlivňující náplň dílčího cíle „1“ a nástin jejich řešení jsou následující: RF1: Během zpracování korpusů a korpusových nástrojů se vyskytnou další korpusy obsahující srovnatelná data. Řešení: Korpusy pro český jazyk vznikají v ČR na celkem pěti pracovištích, která udržují těsné kontakty a výsledky výzkumu si vzájemně vyměňují nebo se o nich poměrně obsáhle informují. Navíc je třeba rozlišovat mezi korpusy psanými (textovými) a řečovými. Řečové korpusy vznikají prakticky jen na pracovištích v Plzni, Brně a Liberci, z nichž dvě se na řešení tohoto projektu budou podílet. Navíc vznik jakéhokoli dalšího korpusu je pozitivním jevem, neboť v tomto oboru více než kdekoli jinde platí, že vhodných dat není nikdy dostatek. Tudíž korpusy vytvořené v rámci navrhovaného projektu budou v každém případě využity i dalšími pracovišti. V případě	- 31.12.2007

		cizojazyčných korpusů budou využívány korpusy, které jsou k dispozici v systému ELRA (European Language Resources Association). RF2: Nepodaří se získat dostatek materiálů, resp. mluvčích, pro vytvoření textových, resp. audiokorpusů. Řešení: Tento rizikový faktor nebude mít zřejmě přílišnou váhu, neboť již současný web poskytuje doslova nepřeberné množství textového materiálu, z nichž lze za použití vhodných vyhledávacích metod vybrat dostatečné množství materiálu pro vytvoření korpusu. V případě řečových korpusů nejde ani tak o problém nalezení vhodné množiny dat nebo množiny vhodných mluvčích, nýbrž kritickým faktorem je čas. Pořizování řečových dat a zejména jejich následné zpracování (třídění, anotace, apod.) vyžaduje značné množství času, avšak riziko lze úspěšně odstranit kvalitním managementem projektu. RF3: V průběhu naplňování dílčího cíle projektu se vyskytne komerční software řešící problematiku pořizování korpusů. Řešení: Pokud se nějaký software vyskytne a bude využitelný, nebude díky modularitě předpokládaného programového vybavení příliš obtížné ho do vytvářeného software začlenit. Pravděpodobnost jeho výskytu v dohledné době je však minimální.	
	2	**Dílčí cíl** Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka. **Indikátory dosažení - výsledky dílčího cíle** a) Návrh formalismu pro popis sémantiky na rozsáhlejší doméně, návrh vhodně strukturovaného sémantického popisu dotazů uživatelů, eventuálně vytvoření vlastního hierarchického systému relací mezi lexémy pro zaručení generalizační schopnosti systému. b) Vytvoření ontologií pro aplikaci formalismu popisujícího sémantiku. Jednotlivými výsledky budou: - návrh ontologie, sémantických konceptů – datový formát XML, vytvoření UML modelu, - návrh ohodnocení jednotlivých konceptů vektorem sémantických příznaků, a to jak doménových, tak obecnějšího charakteru, - návrh soustavy vektorů ohodnocení jednotlivých konceptů. c) Vytvoření multilingválního sumarizačního systému včetně rezoluce anafor a komprese souvětí, jeho zakomponování do prostředí pro vyhledávání a vývoj metod ohodnocování jeho kvality, návrh metod disambiguace v multijazykovém prostředí s využitím kontextu, thesauru a pravděpodobnostních metod: - sumarizační systém obohacený o kompresi souvětí, - systém rezoluce anafor a jeho využití při sumarizaci – pro angličtinu bude využit systém GuiTAR, vytvořený na univerzitě Essex (Anglie), pro češtinu bude na základě poznatků získaných na českých pracovištích vytvořen vlastní systém, - metoda hodnocení kvality sumarizátorů na základě LSA. d) Vývoj nových, dokonalejších modelů elektronických dokumentů tak, aby při použití textových klasifikačních algoritmů bylo dosaženo co nejlepších výsledků při rozpoznávání tématu, rozpoznávání spamových emailů, detekci dokumentů se závadným obsahem apod. e) Vytvoření metodologie a nástrojů pro analýzu webových dokumentů. **Prostředky ověření - Forma zpracování a předání výsledku dílčího cíle** Při naplňování tohoto dílčího cíle půjde o vytvoření základního teoretického podpůrného aparátu, bez něhož nebude možné další kroky realizovat. Jediný tento dílčí cíl bude mít charakter spíše základního výzkumu – půjde o vývoj metod, metodologií a formálních modelů pro návrh zamýšleného komunikačního rozhraní, avšak součástí výzkumných prací bude též experimentální implementace a vytvoření softwarových nástrojů pro evaluaci vyvíjených metod a formalismů. Výsledky budou shrnuty do písemných dokumentů a prezentovány téměř výhradně formou publikací na konferencích, v odborných časopisech a v průběžných výzkumných zprávách. **Kritické poedpoklady dosažení dílčího cíle** Rizikové faktory ovlivňující dosažení dílčího cíle „2“ a nástin jejich řešení mohou být následující: RF1: Nepotvrzení či neplatnost výzkumných hypotéz poskytujících základ pro vytvoření formalismů a modelů. Řešení: Plánovaný dílčí cíl zde nestojí na jediné výzkumné hypotéze, nýbrž na teoretickém základu návrhu komunikačních systémů. Využito bude jak dosavadních poznatků z návrhu existujících komunikačních rozhraní a systémů pro interakci člověka s počítačem, tak i poznatků z	- 31.12.2008

		psychologie komunikace a doporučení TC.13 IFIP (for HCI). Základním rizikem proto bude opět časový faktor, který lze výrazně omezit dobrým managementem projektu. RF2: Nedostatečná erudice členů týmu pro vývoj formálních prostředků. Řešení: Tento rizikový faktor nebude mít zřejmě přílišnou váhu, neboť oba participující týmy jsou složeny minimálně z poloviny ze starších zkušených výzkumníků, z nichž někteří se předmětnou oblastí zabývají 25 i více let, z druhé části pak z mladých perspektivních pracovníků, kteří buď vyrostli anebo se podíleli na řešení podobné problematiky a potřebné teoretické základy oboru již získali, zejména v doktorandském studiu.	
	3	**Dílčí cíl** Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. **Indikátory dosažení - výsledky dílčího cíle** a) Implementace uživatelského rozhraní pro hlasovou komunikaci se sémantickým webem –součástí výsledku budou: - implementace LVCSR rozpoznávače, - natrénování akustických a jazykových modelů, - implementace nahrávacího modulu se stochastickým modelem detekce řečového signálu, - implemetace parametrizátoru na bázi MFCC, - návrh a implementace modulu pro akustické modelování založeného na umělých neuronových sítích nebo směsích Gaussových funkcí, - návrh a implementace efektivního dekódovacího algoritmu, který dokáže pracovat s gramatikami a stochastickými jazykovými modely, - programová realizace a ověření funkčních vlastností robustního ASR systému pro inflexní jazyky. b) Systém pro extrakci významu ze spontánních promluv – dílčími kroky k dosažení tohoto dílčího cíle budou: - návrh a realizace optimální řečové databáze, - návrh systému sémantického značkování řečových dat, - báze znalostí umožňující automatizované či automatické značkování spontánních promluv uložených v databázi, - implementace stochastických sémantických gramatik pro automatickou sémantickou analýzu dotazu uživatele, - využití hierarchické ontologie pro tvorbu strukturalizovaného popisu dotazů uživatele a pro zajištění schopnosti zobecňování z natrénovaných dat, - aplikace metod mělkého (shallow) parsingu promluv pro částečnou analýzu dotazů uživatele. c) Vytvoření komfortního uživatelského rozhraní pro práci se sémantickým webem – součástí tohoto dílčího cíle bude: - návrh příslušného dialogového manageru akceptujícího tzv. kombinovanou iniciativu ve vedení dialogu (mixed initiative), - vytvoření robustního systému pro efektivní a časově nenáročné vyhledávání dat v řečové databázi, - vytvoření robustního a spolehlivého modelu sémantické hierarchie a jeho implementace. d) Aplikace a modifikace OWL standardu v českém prostředí. e) Aplikace klasifikačních metod v multijazykovém prostředí. f) Kompletace multilingválního sumarizačního systému včetně rezoluce anafor a komprese souvětí. g) Algoritmy vhodné pro generování itemsetů a n-gramů a ověření jejich úspěšnosti pro klasifikaci textových dokumentů. h) Výchozí algoritmy pro vyvozování nových znalostí z informací získaných z volného textu. i) Prototyp programu pro přiřazování logických formulí větám z volného textu. **Prostředky ověření - Forma zpracování a předání výsledku dílčího cíle** V dílčím cíli „3“ jde o vytvoření souboru programových produktů, které vzniknou implementací teoretických metod a formalismů vytvořených v rámci dílčího cíle „2“. Výsledky budou mít jednoznačně aplikační charakter, i když vesměs půjde jen o experimentální software, bez něhož nelze metody a modely verifikovat. Výsledky však bude možno předat i dalším zájemcům, protože se předpokládá úplná dokumentace vytvořeného programového vybavení. Výsledky budou prezentovány jako balíky experimentálního software a metod, dále budou publikovány na konferencích, v průběžných výzkumných zprávách, případně také zveřejněny formou speciálních letáků, v tisku a uvažuje se též o možnosti předvedení na specializovaných veletrzích a výstavách. **Kritické poedpoklady dosažení dílčího cíle**	- 31.12.2009

		Rizikové faktory ovlivňující dosažení dílčího cíle „3“ a nástin jejich řešení: RF1: V průběhu projektu přestane být o vytvářené přístupové technologie zájem a pracoviště účastnící se na řešení projektu se tak ocitnou bez reálné využitelnosti svých výsledků. Řešení: Současným trendem je naopak příklon k využívání multimediálních a multimodálních dat, ukládání velkých množství dat a informací na běžných počítačových prostředcích, sílí propojování informačních technologií s rozhlasovým a televizním vysíláním, streamovanými médii a mobilními komunikacemi. Nové hardwarové prostředky budou vyžadovat nové technologie přístupu k datům, přičemž preferována bude komunikace v přirozeném jazyce, ať už psanou nebo mluvenou formou. Vyvíjené programové prostředky tento trend jednoznačně podpoří a proto je toto riziko za dobu řešení projektu téměř nulové. RF2: V průběhu řešení projektu se vyskytne komerční software řešící problematiku srovnatelnou s předpokládanými výsledky projektu. Řešení: Komerční řešení využívající přístup k datům na webu prostřednictvím přirozeného jazyka jsou dosud v plenkách a komerční sféra naopak aktivně vyhledává zajímavé práce z akademické sféry. Proto je toto riziko minimální, očekáváme naopak velký zájem z komerční sféry. RF3: Časové faktory ovlivňující zpracování software. Řešení: Při implementaci a programové realizaci metod vyvinutých v rámci dílčího cíle „2“ může dojít k určité časové tísni vlivem nevhodně zvolených implementačních nástrojů, eventuálně nezkušeností některých mladších členů týmu. Riziko je však minimální, neboť řešitelský kolektiv je složen vesměs ze zkušených výzkumníků a mladých pracovníků, kteří již obdobné, i když jednodušší systémy v minulosti vytvářeli a implementovali. Časový faktor lze navíc výrazně ovlivnit dobrým managementem projektu.	
	4	**Dílčí cíl** Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí. **Indikátory dosažení - výsledky dílčího cíle** a) Testování a ověřovací provoz implementovaného hlasového rozhraní – součástí bude - otestování zpracovaného LVCSR rozpoznávacího systému, - ověření funkčních vlastností robustního ASR systému na vhodné množině uživatelů, - otestování vyvinutých metod automatické sémantické analýzy dotazů. b) Ověření funkčních vlastností vytvořených ontologií a hierarchického systému relací mezi lexémy pro zaručení generalizační schopnosti systému analýzy sémantiky, c) Ověření vlastností algoritmů pro klasifikaci a analýzu dat na různých typech dokumentů. d) Otestování a ověření navržených metod na konkrétních typových řešeních, např. na přístupu k webovým stránkám výzkumných a vzdělávacích institucí. e) Vyhodnocení úspěšnosti jednotlivých fází analýzy volného textu od morfologické úrovně až po převod do logických formulí. **Prostředky ověření - Forma zpracování a předání výsledku dílčího cíle** Náplní dílčího cíle „3“ je provedení rozsáhlých testů (tzv. field experiments) vyvinutých metod, metodologií, modelů a vytvořeného souboru programových produktů. Předpokládá se testování produktů na obvyklých třech skupinách uživatelů – v prvním kroku budou vlastnosti systémů a metod prověřovány úzkou skupinkou řešitelů projektu, ve druhém kroku bude testovací množina uživatelů vytvořena ze spolupracovníků, kteří však s řešením projektu neměli nic společného a o výsledcích řešení jsou jen velmi kuse informováni, a teprve ve třetím kroku bude systém testován libovolnými uživateli, tzv. „lidmi z ulice“. Zčásti však v tomto kroku budou využiti studenti, kteří všeobecně mají tendenci takové systémy „pokořit“. Výsledky budou kompletně dokumentovány a z vyhodnocení experimentů budou vyvozovány příslušné závěry, tj. systém a jeho části budou průběžně doplňovány, upravovány a opětovně testovány. V závěru budou výsledky testování a ověřovacího provozu publikovány v časopisech, na konferencích a obšírně v závěrečné výzkumné zprávě. **Kritické poedpoklady dosažení dílčího cíle** Rizikové faktory ovlivňující dosažení dílčího cíle „4“ a možná řešení: RF1: V průběhu testů se projeví nedostatky v koncepci systému vedoucí k závažným problémům ve funkci systému. Řešení: Řešitelský tým je složen z odborníků, kteří obdobné, i když jednodušší, systémy již vytvořili a mají z jejich tvorby nezanedbatelné zkušenosti. Tým byl dále doplněn o mladé pracovníky, kteří	- 31.12.2010

		se podíleli na tvorbě řady produktů pro prezentace na webových stránkách a je jim problematika přístupu k webu velmi blízká. Riziko volby nevhodné koncepce je proto minimální. RF2: V průběhu testů se projeví nedostatky v implementaci systému a metod. Řešení: Obdobné jako předchozí rizikový faktor – řešitelský tým je složen z odborníků, kteří obdobné, systémy již vytvořili a mají i z jejich implementace poměrně rozsáhlé zkušenosti. Riziko závažných implementačních chyb je proto minimální, drobné nedostatky v implementaci bývají zpravidla v krátké době snadno odstranitelné. RF3: Nepodaří se vytvořit dostatečně reprezentativní množiny testovacích osob. Řešení: Ve vztahu k odstavci 3.3.3. (tři úrovně testování) je riziko nedostatečného vytvoření skupin testujících osob nepatrné – obě participující pracoviště jsou poměrně rozsáhlá a množinu osob testujících vlastnosti systému nebude problém vytvořit; ostatně bylo již ověřeno v minulosti na jednodušších úlohách. Otázka volby třetí skupiny osob je spíše otázkou vytvořeného přístupu k systému – zde se nabízejí dvě možnosti: Buď si osoby vhodné k testování systému vybírat podle určitých hledisek (bylo tak někdy postupováno v minulosti a osoby byly k testování zvány na řešitelské pracoviště) nebo zveřejnit přístupový portál systému a dovolit testování systému široké veřejnosti prostřednictvím internetu, popř. přes telefon (telefonní přístup je však v současných podmínkách omezen kvalitou spojení v mobilních sítích, resp. kvalita spojení je dána úrovní signálu v místech, kde se potenciální uživatel právě nachází, a výsledky testů jím mohou být zkresleny). Rizikový faktor může být opět minimalizován vhodnými rozhodnutími, resp. dobrým managementem projektu.	

2.2.1. AKTIVITY USKUTEČNĚNÉ v roce 2009

Číslo aktivity

01/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Vývoj rozpoznávače JLASER

Zahájení aktivity

5.1.2009

Ukončení aktivity

22.12.2009

Popis aktivity

Automatický rozpoznávač řeči JLASER byl doplněn o další moduly nezbytné pro jeho kompletní funkci. Byla přidána podpora pro parametrizaci PLP (Perceptual Linear Prediction), která má robustnější chování při nepříznivých nahrávacích podmínkách (šum, dozvuk...). Parametrizace v současnosti zohledňuje 3 základní faktory z psychofyziky slyšení: kritické pásmo spektrální citlivosti, křivky stejné hlasitosti a vztah mezi intenzitou a vnímanou hlasitostí. Dosaženým výsledkem jsou koeficienty popisující vyhlazený tvar spektra řečového signálu. Kromě implementace PLP parametrizace byla provedena oprava několika drobnějších programových chyb zjištěných v kódu rozpoznávače při jeho důkladném testování.

Skutečné Indikátory dosažení - výsledky aktivity

Bylo provedeno rozsáhlé testování zpracovaného programového systému a vyhodnoceny testy úspěšnosti rozpoznávání.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Zdrojové kódy programového řešení byly umístěny na webovou stránku projektu. Podrobnější popis výsledků aktivity lze nalézt v přiložené publikaci: Brychcín, T.: PLP parametrizace pro ASR systém JLASER, Semestrální práce z předmětů KIV/AZS a KIV/TKS, KIV ZČU Plzeň, únor 2009.

Číslo aktivity

02/09

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

Poměrové statistiky

Zahájení aktivity

5.1.2009

Ukončení aktivity

22.12.2009

Popis aktivity

Při testování úspěšnosti rozpoznávání je často problém s nedostatkem dat. Pokud jsou např. výsledky dvou testů blízké, musí se vzít v úvahu náhodná chyba a je třeba testovat statistickou významnost. Existující statistické testy většinou předpokládají nezávislost sledovaných jevů (v našem případě je sledovaným jevem rozpoznání slova). Při rozpoznávání vět může chyba v jednom slově způsobit chybu v dalších slovech a proto při rozpoznávání jednotlivých slov může vzniknout statistická závislost. Cílem této aktivity bylo navrhnout metody pro porovnání výsledků rozpoznávání, které berou v úvahu statistickou závislost úspěšného rozpoznání slov ve větě. Předpokladem použití odvozené metody byla závislost mezi chybami ve slovech v rámci jedné věty a nezávislost chyb v různých větách. Ve spolupráci s katedrou matematiky byla zkoumána náhodná proměnná vyjadřující poměr správně rozpoznaných slov ve větě k celkovému počtu slov ve větě. Podařilo se odvodit distribuční funkci této náhodné proměnné a následně postup pro výpočet tolerančních intervalů. Metoda byla vyzkoušena v praxi při porovnávání výsledků akustických modelů založených na neuronových sítích a směsích Gaussových funkcí.

Skutečné Indikátory dosažení - výsledky aktivity

Byla navržena, programově realizována a ověřena originální metodologie pro porovnání úspěšnosti výsledků rozpoznávání.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Dosažené výsledky byly prezentovány formou článků v odborném tisku a vystoupení na konferencích a seminářích. Jejich seznam je uveden v odstavci 4.1.2.

Číslo aktivity

03/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Trénování akustických a jazykových modelů

Zahájení aktivity

5.1.2009

Ukončení aktivity

22.12.2009

Popis aktivity

V roce 2009 pokračoval výzkum akustických modelů založených na neuronových sítích. Ukázalo se, že použití neuronových sítí může být výhodné, pokud je malý slovník. Experimenty byly prováděny na dostupných českých korpusech. Nově bylo provedeno trénování akustických modelů na volně šiřitelném anglickém korpusu VoxForge (www.voxforge.org). Nyní jsou k dispozici natrénované akustické modely pro češtinu i angličtinu, a to jak pro originální recognizer JLASER, tak pro recognizer koncipovaný na bázi HTK. Rovněž byly provedeny testy s trigramovými jazykovými modely a LVCSR rozpoznávačem HDecode. Součástí výzkumu bylo také ověření využitelnosti akustických modelů založených na slabikách. Byly provedeny testy na korpusu spontánní řeči LACS a byly provedeny první pokusy o clustering slabikových jazykových modelů. Z předběžných testů vyplývá, že se slabikovými akustickými modely bude dosahováno lepší úspěšnosti rozpoznávání než při použití triphonových akustických modelů.

Skutečné Indikátory dosažení - výsledky aktivity

Natrénované akustické a jazykové modely.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Byly průběžně vyhodnocovány testy úspěšnosti rozpoznávání, výsledky testů a související publikace jsou uvedeny v odstavci 4.1.2.

Číslo aktivity

04/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Vývoj pokročilého algoritmu pro automatickou sémantickou analýzu dotazů

Zahájení aktivity

5.1.2009

Ukončení aktivity

22.12.2009

Popis aktivity

Tato aktivita navazovala na aktivitu 2008-03 (Vývoj základního algoritmu pro automatickou sémantickou analýzu dotazů) a v jejím rámci jsme se snažili o vylepšení stávajících algoritmů tak, že v současnosti dosahují vyšší přesnosti analýzy vět. Zvýšení přesnosti bylo dosaženo použitím stochastického modelu a vytvořením nového algoritmu pro parsování vět. Nově navržený algoritmus využívá kontextové informace pro zpřesnění statistických

odhadů. V rámci aktivity byly také testovány metody zpracování jazyka specifické pro inflexní jazyky tyto metody však nepřinesly výrazné zvýšení přesnosti parsování.

Skutečné Indikátory dosažení - výsledky aktivity

Indikátorem dosažení výsledku je vyšší přesnost analýzy vět měřená procentem shody s testovací kolekcí vět. Nejlepší algoritmus parsování s použitím kontextu dosahuje 82% shody. Algoritmus, který používá stochastický model, dosahuje 74% úspěšnosti analýzy. Jedná se o zlepšení o zhruba 20%, respektive o 12% oproti základnímu (původně použitému) algoritmu.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledky vývoje algoritmu byly publikovány v odborné literatuře (viz odst. 4.1.2). Programové produkty byly uloženy v datovém úložišti projektu.

Číslo aktivity

05/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Extrakce faktografických dat z veřejně dostupných zdrojů

Zahájení aktivity

5.1.2009

Ukončení aktivity

22.12.2009

Popis aktivity

V rámci této aktivity byly vytvořeny dva experimentální systémy (systém QAS a systém LASE), které jsou schopny na základě dotazu formulovaného v přirozené řeči poskytovat relevantní odpovědi. Nejprve bylo nutné analyzovat strukturu typických dotazů formulovaných v přirozeném jazyce. K tomuto účelu bylo využito dat získaných v rámci aktivit 2007-37 (Pořizování korpusu dotazů z reálného prostředí) a 2008-01 (Sémantické anotování korpusu). Analýza potvrdila, že je vhodné používat dva přístupy k analýze otázek a hledání odpovědí. Pro první přístup (použitý v systému QAS) byly využity výsledky aktivity 2008-03 (Vývoj základního algoritmu pro automatickou sémantickou analýzu dotazů), při druhém přístupu (systém LASE) byl použit systém generických šablon využívající lexikální analýzu.

Skutečné Indikátory dosažení - výsledky aktivity

Oba vytvořené systémy se v současnosti nacházejí v experimentálním stadiu vývoje. Dosavadní vývoj zatím potvrdil, že se jedná o velmi slibné přístupy, které by mohly v brzké době umožnit komunikaci se sémantickým vyhledávacím strojem přirozenou řečí. Systém LASE dosahoval úspěšnosti 72% na testovací množině dotazů, systém QAS dosahoval 81% úspěšnosti.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Systémy jsou k dispozici na adrese
http://liks.fav.zcu.cz/mediawiki/index.php/Research.
Zdrojové kódy a data se nacházejí v repozitáři na adrese https://liks.fav.zcu.cz/svnsemantic/. Struktura a hlavní myšlenky systémů jsou popsány v odborných textech:
Konopík, M.: Hybrid Semantic Analysis, Doctoral Thesis, ZČU Plzeň, 2009.
Saleh, S. T.: Odpovídání dotazů - získání odpovědi z textů, Bachelor Thesis, ZČU Plzeň, 2009.
Detailní popis výsledků je uveden v odstavci 4.1.2.

Číslo aktivity

06/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Ontologie, aplikace OWL a uživatelské rozhraní v oblasti ERP (Event related potentials)

Zahájení aktivity

1.10.2008

Ukončení aktivity

22.12.2009

Popis aktivity

V rámci této aktivity byly aplikovány standardy RDF a OWL v oblasti ERP (evokované potenciály). Byla definována metadata a vytvořena ontologie dané domény. Transformační mechanismus pro výměnu dat a metadat mezi relační databází (objektovým modelem databáze) a technologiemi sémantického webu umožňuje základní trasformaci mezi těmito popisy. Vzhledem k rozdílným vyjadřovacím možnostem relační databáze a jazyka OWL (podstatně větší složitost problému) bude v aktivitě pokračováno i v příštím roce.

Skutečné Indikátory dosažení - výsledky aktivity

Byla realizována ontologie ERP domény a základní transformační mechanismus pro výměnu dat a metadat s relační databází.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledky a vytvořené softwarové nástroje jsou popsány ve studentských (bakalářských) pracích:
Štěbeták, J.: ERP portál a sémantický web - datová vrstva, Bakalářská práce, Plzeň, 2009.
Brůha, P.: ERP portál a sémantický web - aplikační a prezentační vrstva, Bakalářská práce, Plzeň, 2009.

Číslo aktivity

07/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Automatické rozpoznávání dialogových aktů na webových stránkách

Zahájení aktivity

5.1.2009

Ukončení aktivity

22.12.2009

Popis aktivity

Tato aktivita, navazující na 2006-13, 2007-33 a 2008-15, se zabývala automatickým rozpoznáváním dialogových aktů. Zaměřili jsme se na rozpoznávání dialogových aktů z rozhovorů na webových stránkách. Hlavním cílem aktivity bylo rozšíření dosavadního dialogového korpusu o další rozhovory a tyto pokud možno automaticky anotovat dialogovými akty. Pro segmentaci dialogů a automatické rozpoznávání dialogových aktů byla použita zejména větná interpunkce, informace o změně řečníka apod. Ukázalo se, že tyto informace jsou pro rozpoznávání dialogových promluv dostatečné, proto nebyly dříve popsané a implementované metody rozpoznávání zatím použity.

Skutečné Indikátory dosažení - výsledky aktivity

Byl vytvořen nástroj, který po zadání URL prohledá příslušnou stránku, pokusí se nalézt rozhovory, následně segmentuje text na větné jednotky a rozpozná dialogové akty. Nalezené rozhovory s anotovanými dialogovými akty jsou uloženy do databáze a zároveň je k dispozici výstup v podobě XML souborů.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Nástroj byl ověřen na rozhovorech v internetovém deníku Super (http://www.super.cz/svet-celebrit/rozhovory/). Bylo detekováno a zpracováno 382 rozhovorů. Tyto rozhovory byly automaticky anotovány dialogovými akty. Celkem bylo anotováno 31058 dialogových aktů. Anotované rozhovory jsou k dispozici na http://home.zcu.cz/~pkral/dialogs.zip. Související publikace jsou uvedeny v detailním popisu výsledků (viz odstavec 4.1.2) a v souhrnném seznamu publikací.

Číslo aktivity
08/09
Ke kterému dílčímu cíli se aktivita vztahuje
3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...
Název (cíl)aktivity
Kategorizace dokumentů neuronovou sítí
Zahájení aktivity
5.1.2009
Ukončení aktivity
22.12.2009
Popis aktivity
V počáteční fázi aktivity byly dokončeny rozsáhlé testy zaměřené na optimální nastavení parametrů sítě ART-2 používané pro kategorizaci dokumentů. V další fázi se naše činnost zaměřila na návrh a implementaci neuronové sítě učené s učitelem, vhodné pro kategorizaci dokumentů. V tomto případě se jednalo o síť typu vícevrstvý perceptron (MLP). Síť byla zařazena jako druhý stupeň hierarchického modelu kategorizátoru. Jako první stupeň modelu (kategorizátor slov) bylo i nadále využíváno Kohonenovy mapy (SOM). Navržený model byl implementován a natrénován na dostatečném množství vstupních dokumentů (tiskové zprávy ČTK). Výsledky kategorizace byly porovnány s modely SOM a ART-2 a jsou podrobně shrnuty v bakalářské práci M. Valenty. Vzhledem k tomu, že výsledky kategorizace jsou do značné míry ovlivněny transformací vstupního textu do číselné podoby (vytvoření tzv. kontextového vektoru), byla zvažována i možnost náhrady transformačního algoritmu algoritmem vhodnějším pro zpracování česky psaných dokumentů. Vzhledem k syntaxi češtiny zatím nepodařilo navrhnout "lepší" způsob reprezentace vstupního vektoru a tak se i nadále bude jako vstupní vektor kategorizátoru slov používat n-gramový model.
Skutečné Indikátory dosažení - výsledky aktivity
Byla implementována neuronová síť MLP a začleněna do programového systému pro kategorizaci dokumentů. Byly provedeny základní testy kategorizace dokumentů s touto sítí.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity
Výsledky a softwarové nástroje byly popsány v pracích:
Valenta, M.: Aplikace neuronových sítí v oblasti zpracování česky psaných dokumentů, Bakalářská práce, Plzeň, 2009.
Mouček, R., Mautner, P.: WEBSOM method - word categories in Czech written documents, In: Proceedings of Int. Conf. Text, speech and dialogue 2009. Springer Verlag, Berlin, Heidelberg, 2009, s. 85-92. ISBN 978-3-642-04207-2.

Číslo aktivity
09/09
Ke kterému dílčímu cíli se aktivita vztahuje
3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...
Název (cíl)aktivity
Zpracování korpusových záznamů (korpusy LAC-SS a LAC-Noise)
Zahájení aktivity
5.1.2009
Ukončení aktivity
22.12.2009
Popis aktivity

V průběhu období byla navržena metodika zpracování a katalogizace záznamů v korpusech LAC-SS 2007 a 2008, jejich úpravy, předzpracování, transkripce a příprava k využití pro trénování akustických modelů ASR systému JLASER, k přípravě databáze segmentů pro TTS systém jSynt (Ing. Pavel Král, Ph.D.) a k návrhu a ověřování metod předzpracování akustického signálu pro potřeby ASR. Korpusy byly ve spolupráci s databázovou skupinou (Ing. M. Zíma, Ph.D.) uloženy do zabezpečeného úložiště v souborovém systému AFS. Korpusy (ze zjevných důvodů) nelze považovat za uzavřené a dokončené, sběr materiálu pokračuje při každé vhodné příležitosti... Nadále pokračoval sběr dat do ruchového korpusu LAC-Noise, katalogizace a úpravy těchto záznamů byla provedena řada testů v oblasti algoritmů adaptivní fitrace řečového signálu s cílem zvýšení výkonu ASR systému odstraněním pozaďových ruchů a šumu. Během těchto testů se ukázalo jako vhodné pořizovat šumový signál ve vyšší kvalitě, ideálně stereo párem shodných vysoce citlivých mikrofonů na rozlehlé bázi, neboť informace o fázi šumového signálu lze také při filtraci s úspěchem využít. Nasbírané záznamy byly použity při návrhu a ověřování metod zvýšení výkonu a spolehlivosti ASR systémů v obtížných příjmových podmínkách a také při zkoumání psychoakustických parametrů vnímání akustického signálu člověkem. Navržené metody adaptivní filtrace jsou momentálně ve stádiu přípravy k zapracování do ASR systému JLASER.

Skutečné Indikátory dosažení - výsledky aktivity

- V úložišti jsou kvalitně připravené korpusy LAC-SS 2007, LAC-SS 2008 a částečně LAC-SS 2009, určené k použití při výše zmíněných aktivitách, opatřené transkripcí, příp. anotací, bezpečně uložené, dobře organizované a snadno systematicky přístupné. Pro práci s korpusy byly kvalitně vyškoleny dvě pracovnice na přepis korpusových záznamů (transkripci).

K dispozici jim byl sepsán anotační manuál (aktualizovaná verze), který pokrývá většinu situací, se kterými se mohou při transkripci setkat.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

- Korpusy spontánní řeči LAC-SS
- Autorizovaný software LTranscriber (http://www.kiv.zcu.cz/vyzkum/software/detail.html?id=54) vyvinutý jako bakalářská práce studenta M. Konkola pod vedením Ing. K. Ekšteina, Ph.D. a dalších členů týmu (viz odst. 4.1.2 a souhrnný seznam publikací).

Číslo aktivity

10/09

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

Ověření a testy modulů pro zpracování přirozeného zpracování jazyka v multilinguálním prostředí.

Zahájení aktivity

4.1.2009

Ukončení aktivity

9.7.2009

Popis aktivity

Aktivita byla pokračováním aktivity č. 10 z roku 2008. V rámci této aktivity byly testovány a ověřeny výsledky jednotlivých modulů pro zpracování přirozeného jazyka ve vícejazyčném prostředí. Srovnán a vyhodnocen byl vliv multilinguality textových dat na výsledky jednotlivých metod zpracování přirozeného jazyka. Důraz byl kladen na srovnání možností předzpracování textu do jazykově nezávislé formy. Navíc oproti plánu byl zdokonalen a otestován disambiguační modul.

Skutečné Indikátory dosažení - výsledky aktivity

Implementovaný prototypový systém byl testován na předzpracované kolekci článků agentur Reuters a ČTK převedené do jazykově nezávislé formy. Byla sledována především relevance a úplnost vyhledávání. Vypracování této aktivity bylo oproti plánu rozšířeno o testy disambiguačního modulu, který byl zkoušen na standardním korpusu SensEval2 - Semcor, kde se algoritmus umístil na 2. místě z hlediska míry F1 ze 24 zúčastněných algoritmů.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Systém byl otestován na relevanci získaných výsledků jak pro české a anglické prostředí, tak i při křížovém

zpracování. Ověřeny byly také komparativní výsledky získané při srovnání s vyhledávačem Google Desktop Search. Systém jsme dále evaluovali na korpusech textů Evropského parlamentu. Aktivita vedla k ověření výstupů autorizovaného softwaru uvedeného v aktivitě 11/09.

Číslo aktivity

11/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Implementace systému pro vyhledávání a sumarizaci

Zahájení aktivity

5.1.2009

Ukončení aktivity

30.6.2009

Popis aktivity

Aktivita navazovala na výstupy z návrhu systému MUSE pro vyhledávání dokumentů v multilingválním prostředí. Prototypový systém byl implementován včetně zamýšlených optimalizací.

Skutečné Indikátory dosažení - výsledky aktivity

Byl vytvořen systém umožňující vyhledávání nad datovým korpusem vytvořeným z tiskových zpráv Reuters a ČTK, který obecně může obsahovat jakákoliv textová data. Díky převedení textu do jazykově nezávislé formy je možné křížové zpracování. Systém umožňuje vyhledávání na základě dotazu jak v plných textech, tak v extraktech, což výrazně zvyšuje přehlednost se zachováním téměř shodné relevance výsledků dotazu. Prototypový systém je vytvořen z několika hlavních modulů, tedy je možné měnit a vylepšovat jeho součásti nezávisle na zbytku.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledkem aktivity je programový systém zveřejněný formou autorizovaného softwaru včetně potřebné uživatelské dokumentace a ukázkových dat - viz http://www.kiv.zcu.cz/vyzkum/software/.

Číslo aktivity

12/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Metoda aktualizační sumarizace

Zahájení aktivity

5.1.2009

Ukončení aktivity

30.10.2009

Popis aktivity

V předchozí etapě projektu byla zkoumána sumarizace tématu. Cílem této aktivity bylo rozšířit sumarizační metodu tak, aby se ve výsledném souhrnu nevyskytovaly základní/redundantní informace. Souhrn pak obsahuje pouze nové informace – aktualizační sumarizace (Update Summarization). Na vstupu byly dva shluky dokumentů pojednávajících o stejném tématu/události: dokumenty v prvním shluku uživatel již četl, dokumenty ve druhém ještě ne. Cílem bylo vytvořit souhrn „nových“ dokumentů, kde se nevyskytují informace uživateli již známé. V rámci aktivity byla vytvořena metoda založená na LSA/IRR. Dle ní vytvořený systém byl použit pro generování souhrnů pro TAC’08 a TAC’09. Na TAC experimentech jsme dosáhli pozoruhodných výsledků: TAC’08: základní souhrny-9/58 (Pyramid score) aktualizační souhrny-7/58 (Pyramid score) TAC’09: základní souhrny-2/52 (Responsiveness) aktualizační souhrny-8/52 (Pyramid score) Pro TAC’09 jsme metodu vylepšili o využití informací o entitách v textech.

Spolupracovali jsme s Universita di Trento (M. Poesio) a JRC, Ispra (R. Steinberger, M. Kabadjov, B. Pouliquen). Práce na tomto tématu byla publikována na TAC'09, TSD'09, ACM DocEng'09, IAPWNC'09 (viz odst. 4.1.2).

Skutečné Indikátory dosažení - výsledky aktivity

V rámci řešení byl vytvořen experimentální systém, který vytváří jak základní souhrny, tak i aktualizační souhrny (autorizovaný software - viz dále). Metoda byla publikována ve sbornících konferencí TAC'08 (Zpráva o '08 TAC experimentech), TSD'09 (metoda založená na LSA/SVD), ACM DocEng'09 (metoda založená na LSA/IRR), IAPWNC'09 (rozšíření o informace o entitách). Dále byl publikován článek o hodnocení kvality sumarizace v časopise CAI detailní citace všech publikací je uvedena v podrobném popisu aktivity v odstavci 4.1.2.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Ověření kvality implementace bylo provedeno v rámci účasti na TAC (Text Analysis Conference - NIST). Výsledky vyvinutého sumarizátoru byly porovnány s výsledky ostatních skupin, které se v rámci konference zúčastnily experimentů.

Číslo aktivity

13/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Online vyhledávací a sumarizační systém

Zahájení aktivity

5.1.2009

Ukončení aktivity

22.12.2009

Popis aktivity

Cílem této aktivity bylo rozšíření online sumarizačního systému SWEeT (http://tmrg.kiv.zcu.cz:8080/sweet) o možnost vytváření aktualizačních souhrnů. Systém pracuje následovně: Uživatel vloží dotaz, který by měl být dostatečně bohatý, aby vymezil dané téma. Navíc zadá datum, které bude oddělovat redundantní informace (starší dokumenty, s jejichž obsahem by měl být již seznámen) a nové informace (z dokumentů, jejichž datum je větší než zadané). Vyhledané dokumenty pak zpracuje zabudovaný systém aktualizační sumarizace a výsledný souhrn, který by měl obsahovat pouze nové informace o daném tématu, bude vrácen uživateli. Pro vlastní sumarizaci byl použit systém ALMUS. Metoda je detailně popsána v příspěvku na konferenci TSD'09 (viz odst. 4.1.2).

Skutečné Indikátory dosažení - výsledky aktivity

Byl vytvořen on-line vyhledávací a sumarizační systém. Jeho architektura a popis jednotlivých částí lze najít v přiložené publikaci. Dále byl systém rozšířen o aktualizační sumarizaci. Rozšířený systém je dostupný na:
http://tmrg.kiv.zcu.cz:8080/usweet/

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Systém byl otestován několika uživateli, byly zaznamenávány jejich dotazy a odpovědi systému. V závěru byla anotována lexikální a obsahová kvalita souhrnu.

Číslo aktivity

14/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Návrh a implementace pokročilé metody pro odhalování plagiátů s využitím LSA

Zahájení aktivity

1.7.2009

Ukončení aktivity

22.12.2009

Popis aktivity

Aktivita volně navazovala na aktivitu č. 18 z roku 2008. V rámci této aktivity byla navržena a implementována pokročilejší metoda pro odhalování plagiátů založená na latentní sémantické analýze. Specifickými oblastmi byly techniky předzpracování textu a jejich vliv na přesnost odhalování plagiátů. Kromě jiného byla prozkoumána kompresní technika založená na náhodném indexování, jejímž cílem je redukce příznaků a snížení časových požadavků. Navržená metoda byla implementována a ověřena na kolekci 1500 českých plagiovaných textových dokumentů.

Skutečné Indikátory dosažení - výsledky aktivity

Implementovaná metoda v podobě DLL knihovny v jazyce C# (.NET Framework 3.5). Ověření funkcionality bylo provedeno na označkovaném korpusu českých plagiátů z aktivity č. 16 z roku 2008. Detailní popis metody lze nalézt v: Češka, Z.: Automatic Plagiarism Detection Based on Latent Semantic Analysis. PhD Thesis, Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic, August 2009. Dalšími výsledky aktivity je řada publikací na vědeckých konferencích (viz odst. 4.1.2).

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Navržená metoda byla odladěna a testována na relevanci výsledků s označkovaným českým korpusem z aktivity č. 16 z roku 2008. Implementovaná metoda v podobě DLL knihovny je k dispozici na webových stránkách http://textmining.zcu.cz/public/NPV/2009/aktivita14/SVDPlag.zip.

Číslo aktivity

15/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Transformace zvolené ontologie do logického programu v jazyce Datalog

Zahájení aktivity

5.1.2009

Ukončení aktivity

22.12.2009

Popis aktivity

Během formulace této aktivity bylo předpokládáno, že pro každou ontologii bude nutné navrhnout nový postup transformace do logického programu. Jak je uvedeno v řešení aktivity 2009-16, byl přímo navrhnut postup, jak téměř z libovolného XML dokumentu vytvořit odpovídající logický program v jazyce Datalog. Protože ontologie jsou zapisovány v jazycích typu RDF, RDFS a OWL, které jsou založeny na jazyce XML, je možné navržený postup na tyto ontologie aplikovat. Datová kolekce studijních programů, oborů a předmětů vysokých škol a univerzit v ČR, která byla vytvořena v rámci aktivity 2008-12 a je zapsána v XML formátu, nevyžaduje žádnou ontologii pro vyhodnocení typických dotazů (např. jak se liší skladba vybraného studijního programu na různých školách). Zodpovězení tohoto dotazu již umožňuje logický program, který získáme transformací dané datové kolekce. Výsledný logický program je nutné doplnit o definici predikátu dotazu v podobě jednoho či více pravidel. Dále byla v rámci jedné diplomové práce zdokonalena „ontologie katastrof“ (aktivita 13-2008) a také přepracována příslušná webová aplikace sémantického vyhledávání. Pro zápis ontologie byl v tomto případě použit jazyk založený na principu trojic (objekt, predikát, předmět).

Skutečné Indikátory dosažení - výsledky aktivity

Způsobem uvedeným v aktivitě 2009-16 byl navržen postup vytvoření logického programu v jazyce Datalog. Ověření správnosti postupu bylo provedeno na fragmentu datové kolekce studijních programů, oborů a předmětů vysokých škol a univerzit v ČR. Z důvodu ručního zpracování a rozsahu datové kolekce byly vzaty v úvahu 2 školy (z každé pouze jedna fakulta, jeden studijní program a obor). Transformaci většího množství dat umožní programový nástroj, který by měl být vytvořen v rámci aktivity roku 2010.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledný logický program napsaný v jazyce Datalog, který vznikl transformací fragmentu zmíněné datové kolekce, je

součástí níže uvedené technické zprávy. Proces vytvoření ontologie katastrof a využití této ontologie k sémantickému vyhledávání je součástí níže uvedené diplomové práce.

Dostal, M.: Zodpovídání dotazů v prostředí sémantického webu. Diplomová práce, Západočeská univerzita v Plzni, 2009.

Zíma, M.: Transformace XML dokumentu do podoby logického programu. Technická zpráva č. DCSE/TR-2009-13, Západočeská univerzita v Plzni, 2009.

Číslo aktivity

16/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Zobecnění transformace ontologie do logického programu v jazyce Datalog

Zahájení aktivity

5.1.2009

Ukončení aktivity

22.12.2009

Popis aktivity

Hlavním cílem této aktivity mělo být zobecnění transformace ontologie do logického programu napsaného v jazyce Datalog. Protože použitá ontologie je většinou zapsána v jazyce, který je založen na jazyce XML, byla řešena obecná problematika transformace XML dokumentu do podoby logického programu. Byl navržen takový postup, ve kterém výsledný logický program obsahuje univerzální pravidla. To znamená, pokud budou tímto způsobem transformovány dva různé XML dokumenty, budou se lišit pouze množinami faktů, množina pravidel bude v obou logických programech shodná. Navržený transformační proces má nevýznamné omezení, které se týká vstupního XML dokumentu, ve kterém je zakázáno, aby značky měly smíšený obsah.

Skutečné Indikátory dosažení - výsledky aktivity

Byl navrhnut postup transformace XML dokumentu do podoby logického programu. Navržený postup byl úspěšně ověřen na testovacích XML dokumentech.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Navržený postup transformace je popsán v technické zprávě:
Zíma, M.: Transformace XML dokumentu do podoby logického programu. Technická zpráva č. DCSE/TR-2009-13, Západočeská univerzita v Plzni, 2009.

Číslo aktivity

17/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Rozšířené možnosti využití on-line slovníku SPOT

Zahájení aktivity

5.1.2009

Ukončení aktivity

22.12.2009

Popis aktivity

Pokračoval provoz webové aplikace slovníku a její základní údržba. Díky vylepšení importu externích korpusů byly

importovány dostupné zdroje překladů od několika autorů. Dále pokračovalo rozšiřování možností slovníkové aplikace pro podporu překladatelských týmů a sofistikovanou práci s taxonomiemi (předdefinované kategorie a uživateli zadávané štítky k pojmům). Vzhledem k nutnému restrukturování implementace, bez něhož by toto rozšiřování vedlo ke zvyšování chybovosti implementace, se oproti původnímu plánu předpokládá ukončení těchto prací v prvním pololetí 2010. Na projektu spolupracují dva studenti navazujícího studia v rámci svých diplomových prací (H.Rysová, T.Brandl – vedoucí P.Brada). Projekt SPOT je mimo využití v rámci projektu dále využíván překladatelským portálem Blogspot.cz jako zdroj referenčních překladů.

Skutečné Indikátory dosažení - výsledky aktivity

Vylepšená verze aplikace dostupná na http://spot.zcu.cz/ naplněná několika korpusy. Rutinní provoz s omezenou množinou uživatelů.

Doprovodný blog na adrese http://blogspot.zcu.cz s průměrnou návštěvností 2000 uživatelů za měsíc, a „spin-off" portál http://blogspot.cz/ využívající slovníkovou databázi SPOT.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Funkční aplikace dostupná na http://spot.zcu.cz/ naplněná příslušnými korpusy.

Číslo aktivity

18/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Metody automatického rozpoznávání dialogových aktů pro zpracování multimediálních vstupů

Zahájení aktivity

5.1.2009

Ukončení aktivity

22.12.2009

Popis aktivity

Tato aktivita navazovala na aktivity 2008-15 a 2008-22 a zabývala se automatickým rozpoznáváním dialogových aktů. Na rozdíl od dosavadního přístupu byla provedena integrace paralingvistických atributů dialogu, tzn. vstupem do systému rozpoznávání je videosignál (zvuk i obraz). Takový model rozpoznávání dialogových aktů přispívá značnou měrou ke zvýšení jeho přesnosti. Byla navržena nová metoda automatického rozpoznávání dialogových aktů, která pracuje přesněji než metody existující, a dále pak analýza účinnosti příznaků používaných pro rozpoznávání dialogových aktů. K tomu byl zvolen poněkud netradiční prostředek – datový sklad. Cílem dalšího zkoumání byla hlubší analýza prozodických příznaků a jejich vazba na výraz obličeje. Výraz obličeje nese velmi důležitou informaci o významu promluvy. Pro analýzu prozódie jsou používány základní prozodické příznaky, v případě analýzy výrazu obličeje byla již navržena klasifikace do pěti základních tříd.

Skutečné Indikátory dosažení - výsledky aktivity

Výsledkem aktivity je nová metoda automatického rozpoznávání dialogových aktů využívající multimediálního vstupu a analýza účinnosti prozodických charakteristik. Metody byly zatím prověřovány na existujících souborech dat získaných záznamem interaktivních dialogů v létech 2007-08, do budoucna se počítá se záznamem dalších (alternativních) forem vedení dialogů.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Navržená metoda a výsledky její analýzy byly prezentovány na významné konferenci (Klečková, J.: Important Nonverbal Attributes for Spontaneous Speech Recognition. In: ICONS´09, March 2009, Cancun, Mexico).

Číslo aktivity

19/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity
Úpravy a využití lexikální databáze VerbaLex obsahující valenční rámce českých sloves v algoritmech syntaktické a logické analýzy
Zahájení aktivity
1.1.2009
Ukončení aktivity
31.12.2009
Popis aktivity
Komplexní valenční rámce v databázi VerbaLex poskytují informace pro pokročilou syntaktickou a logickou analýzu české věty. V rámci dané aktivity byly navrženy a ověřovány konkrétní metody a algoritmy pro jejich využití v systému synt vyvíjenému v Centru ZPJ.
Skutečné Indikátory dosažení - výsledky aktivity
Byla vytvořena nová verze lexikální databáze VerbaLex.

V rámci aktivity byla rozšířena funkcionalita syntaktického analyzátoru synt o možnost získat optimální dekompozici věty na vybrané syntaktické struktury/fráze. Syntaktický analyzátor synt byl využit pro změření pokrytí valenčního slovníku Verbalex.
Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity
Pala, Karel - Horák, Aleš. Multilingual Features of Complex Valency Frames. In Recent Advances in Intelligent Information Systems. Warsaw, Poland : Academic Publishing House EXIT, 2009. od s. 41-49, 9 s. ISBN 978-83-60434-59-8.

Grác, Marek. Shallow Ontology Based on VerbaLex. In NLP, Corpus Linguistics Corpus Based Grammar Research. Bratislava, SR : Slovenská akadémia vied, Jazykovedný ústav Ľ. Štúra, 2009. 400 s. ISBN 978-80-7399-875-2

Jakubíček, Miloš - Kovář, Vojtěch - Horák, Aleš. Measuring Coverage of a Valency Lexicon using Full Syntactic Analysis. In RASLAN 2009 : Recent Advances in Slavonic Natural Language Processing. 1. vyd. Brno : Masaryk University, 2009. od s. 75-79, 5 s. ISBN 978-80-210-5048-8.

Číslo aktivity
20/09
Ke kterému dílčímu cíli se aktivita vztahuje
3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...
Název (cíl)aktivity
Návrh a vývoj nástrojů DEB platformy
Zahájení aktivity
1.1.2009
Ukončení aktivity
31.12.2009
Popis aktivity
Nástroje platformy Dictionary Editor and Browser (DEB) mají široké využití, kde aktuálně vyvíjené systémy zahrnují aplikace pro Global WordNet Grid (vícejazyčná sémantická síť) a multiliguálně orientovanou lexikografickou stanici.
Skutečné Indikátory dosažení - výsledky aktivity
Implementované rozšířené a nové nástroje a metody na platformě DEB. V uvedené aktivitě došlo k inovaci a rozšíření zejména nástrojů DEBDict, DEBVisDic, Global Wordnet Grid,
TeDi nebo Praled. Všechny uvedené nástroje jsou dostupné na adrese http://deb.fi.muni.cz/.

Aplikace DEBVisDic je, mimo jiné, použita jako součást řešení evropského projektu KYOTO (Knowledge Yielding

Ontologies for Transition-based Organization, EU FP7 ICT-211423). Využívá se pro ukládání a úpravy sémantických sítí v různých jazycích. Nově navržené a vyvíjené uživatelské rozhraní vyžaduje návrh a implementaci nových funkcí do rozhraní DEBVisDic API - konkrétní popis viz publikace.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Horák, Aleš - Rambousek, Adam - Vossen, Piek - Segers, Roxane - Maks, Isa - van der Vliet, Hennie. Cornetto Tools and Methodology for Interlinking Lexical Units, Synsets and Ontology. In Current Issues in Unity and Diversity of Languages. Seoul, Republic of Korea : The Linguistic Society of Korea, 2009. od s. 2695-2713, 19 s. ISBN 978-89-90696-71-7.

Horák, Aleš - Rambousek, Adam. Using Wordnets and Ontologies for Text-Meaning Assignment - Implementation Details of the KYOTO Project First Phase. In Proceedings of the 4th International Conference on Software and Data Technologies, Volume 2. Portugalsko : INSTICC, 2009. od s. 303-307, 5 s. ISBN 978-989-674-010-8.

Číslo aktivity

21/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Aplikace syntaktické analýzy pro určení syntaktických struktur ve velmi velkých korpusech

Zahájení aktivity

1.1.2009

Ukončení aktivity

31.12.2009

Popis aktivity

Vyvíjený syntaktický analyzátor s využitím budovaného korpusu syntaktických stromů poskytuje na českých větách strukturní informace vyšší úrovně, které výrazně podporují inteligentní analýzu rozsáhlých textů. V rámci dané aktivity byl systém syntaktické analýzy aplikován, testován a vylepšován vůči vytvářenému velmi velkému (stovky milionů pozic) českému korpusu.

Skutečné Indikátory dosažení - výsledky aktivity

V rámci vyvíjených syntaktických analyzátorů byla implementována extrakce alternativních informací ze syntaktické analýzy: fráze, kolokace a slovesné valence.

S využitím těchto vlastností bylo provedeno základní měření pokrytí lexikonu slovesných valencí na korpusových datech.

Syntaktická analýza byla dále aplikována na sestavení kolokačních statistik (word sketches) pro velké české korpusy.

Pokračovaly práce na vývoji syntaktické analýzy metodou přebíjejících se vzorů (syntaktický analyzátor SET: http://nlp.fi.muni.cz/projekty/set/).

Získaný analyzátor byl vyhodnocen proti dostupným syntakticky anotovaným korpusovým datům. Byla též provedena hrubá analýza kvality anotace těchto dat a identifikace základních problémů v anotaci.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Horák, Aleš - Rychlý, Pavel - Kilgarriff, Adam. Czech Word Sketch Relations with Full Syntax Parser. In After Half a Century of Slavonic Natural Language Processing. Brno, Czech Republic : Masaryk University, 2009. od s. 101-112, 12 s. ISBN 978-80-7399-815-8.

Jakubíček, Miloš - Horák, Aleš - Kovář, Vojtěch. Mining Phrases from Syntactic Analysis. In Text, Speech, Dialogue

2009. 1. vyd. Berlin Heidelberg : Springer Verlag, 2009. od s. 124-130, 7 s. ISBN 978-3-642-04207-2.

Horák, Aleš - Rychlý, Pavel. Discovering Grammatical Relations in Czech Sentences. In Proceedings of the RASLAN Workshop 2009. první. Brno : Masaryk University, 2009. od s. 81-90, 9 s. ISBN 978-80-210-5048-8.

Jakubíček, Miloš - Kovář, Vojtěch - Horák, Aleš. Measuring Coverage of a Valency Lexicon using Full Syntactic Analysis. In RASLAN 2009 : Recent Advances in Slavonic Natural Language Processing. 1. vyd. Brno : Masaryk University, 2009. od s. 75-79, 5 s. ISBN 978-80-210-5048-8.

Kovář, Vojtěch - Jakubíček, Miloš. Prague Dependency Treebank Annotation Errors: A Preliminary Analysis. In RASLAN 2009 : Recent Advances in Slavonic Natural Language Processing. 1. vyd. Brno : Masaryk University, 2009. od s. 101-108, 8 s. ISBN 978-80-210-5048-8.

Kovář, Vojtěch - Horák, Aleš - Jakubíček, Miloš. Syntactic Analysis as Pattern Matching: The SET Parsing System. In Proceedings of 4th Language & Technology Conference. Poznań (Poland) : Wydawnictwo Poznańskie, 2009. od s. 100-104, 5 s. ISBN 978-83-7177-746-2

Číslo aktivity

22/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Rozšíření algoritmů logické analýzy českých vět

Zahájení aktivity

1.1.2009

Ukončení aktivity

31.12.2009

Popis aktivity

Logická analýza věty umožňuje zpracovat základní sémantické znalosti a vztahy v dané větě. V rámci aktivity byla doplňována nová pravidla a navrženy a implementovány nové metody tvorby logické konstrukce české věty na základě její syntaktické analýzy, která využívá budované pokročilé jazykové zdroje, jako jsou komplexní valenční rámce z databáze VerbaLex.

Skutečné Indikátory dosažení - výsledky aktivity

Systém tvorby logické konstrukce v syntaktickém analyzátoru synt v souladu s Principem kompozicionality využívá popisu funkcí pro tvorbu složených logických konstrukcí ze vstupních podkonstrukcí. Předpisy těchto funkcí jsou součástí metagramatiky syntu, kde popisují kompletní tvorbu logické konstrukce na základě výstupních datových struktur syntaktické analýzy. V dané aktivitě byl tento popis rozšířen na cca 180 pravidlových schémat pro tvorbu logické konstrukce - ve výsledku je nyní systém schopen budovat logické konstrukce běžných korpusových vět. Přesnost a komplexní logická správnost výstupních konstrukcí je nyní závislá převážně na úplnosti vstupních informací o lexikálních významech slov - v současné verzi jsou tyto informace založeny na hlavní slovnědruhové klasifikaci a subklasifikaci, což je dostačující sice pro většinu bežných slov, velké množství jevů je ovšem nutné ještě popsat specificky.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Horák, Aleš - Rambousek, Adam. Using Wordnets and Ontologies for Text-Meaning Assignment - Implementation Details of the KYOTO Project First Phase. In Proceedings of the 4th International Conference on Software and Data Technologies, Volume 2. Portugalsko : INSTICC, 2009. od s. 303-307, 5 s. ISBN 978-989-674-010-8.

Kříž, Petr - Logická analýza v rámci systému SYNT. Diplomová práce. Masarykova univerzita, 2010.

Číslo aktivity

23/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Vytvoření velmi velkého českého korpusu

Zahájení aktivity

1.1.2009

Ukončení aktivity

31.12.2009

Popis aktivity

Pro ověřování úspěšnosti jednotlivých metod a algoritmů je vhodné mít co největší množství dat. Proto byl vytvořen velmi velký korpus Czes obsahující české texty. Korpus byl anotován na různých úrovních, aby byl využitelný v co největším množství aplikací.

Skutečné Indikátory dosažení - výsledky aktivity

Korpus Czes má rozsah asi 1,2 miliardy tokenů, je plně označkován na morfologické úrovni pomocí morfologického analyzátoru Ajka s využitím taggeru desamb. V korpusu jsou vyznačeny syntaktické vztahy jednotlivých slov pomocí gramatiky v nástroji Word Sketch Engine.

Daná aktivita má těsnou vazbu na předchozí aktivity 21/09, v níž byly syntaktické vztahy v části korpusu vyznačeny i pomocí syntaktického analyzátoru synt.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Korpus Czes včetně všech značkování je v současnosti k dispozici vybraným uživatelům pomocí webového rozhraní.

Pomikálek, Jan - Rychlý, Pavel - Kilgarriff, Adam. Scaling to Billion-plus Word Corpora. Advances in Computational Linguistics, Mexiko : Instituto Politécnico Nacional, 41, zima 2009, od s. 3-13, 14 s. ISSN 1870-4069. 2009.

Číslo aktivity

24/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Systém pro analýzu anaforických vztahů

Zahájení aktivity

1.1.2009

Ukončení aktivity

31.12.2009

Popis aktivity

Aktivita si klade za všeobecný cíl implementovat program, který automaticky hledá a analyzuje anaforické vztahy ve volných textech.

Skutečné Indikátory dosažení - výsledky aktivity

Probíhalo začleňování dalších zdrojů dat do procesu automatické analýzy anaforických vztahů (AR). S ohledem na praktické použití systému se prioritou ukázalo být propojení se syntaktickým analyzátorem synt, což nově umožňuje vyhledávat anaforické vztahy ve zcela nepředzpracovaných textech. Funkčnost systému byla také doplněna o visualisaci za pomoci systému MMAX2.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Němčík, Václav. The Saara Framework: An Anaphora Resolution System for Czech. In RASLAN 2009 : Recent Advances in Slavonic Natural Language Processing. 1. vyd. Brno : Masaryk University, 2009. od s. 49-54, 6 s. ISBN 978-80-210-5048-8.

Číslo aktivity

25/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Klasifikace a podobnost matematických textů ve vytvořeném korpusu

Zahájení aktivity

1.1.2009

Ukončení aktivity

31.12.2009

Popis aktivity

Aktivita se soustředila na využití shromážděného korpusu klasifikovaných matematických článků a analýzu jazyka matematiky v závislosti na tématice dokumentu. Byly připraveno provedení experimentů s indexováním a vyhledáváním v matematickém korpusu a s~využitím podobnosti a klasifikace pro vyhledávání.

Skutečné Indikátory dosažení - výsledky aktivity

Prototyp aplikace navržené pro nasazení v České
digitální matematické knihovně DML-CZ.
Aplikace pro počítání podobnosti pomocí tří metod (TFIDF, LSI, Random projections) byla implementována a použita na webu http://dml.cz (beta),
a podobnosti spočítány na korpusu více jak 30000 článků.
Získali jsme zpětnou reakci od autorů matematických článků s cílem vybrat nejlépe fungující metriku či vážit vhodně tři existující podobnosti do jedné finální metriky.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

organizace workshopu DML 2009 http://www.fi.muni.cz/~sojka/dml-2009.html a publikace v odborných sbornících:

Sojka, Petr.
DML 2009 Towards a Digital Mathematics Library.
Brno, Czech Republic : Masaryk University Press, 2009. 149 s. ISBN 978-80-210-4781-5.

Sojka, Petr.
An Experience with Building Digital Open Access Repository DML-CZ.
In Proceedings of CASLIN 2009, Institutional Online Repositories and Open Access, 16th International Seminar. první. Pilsen, Czech Republic : University of West Bohemia, Pilsen, 2009. od s. 74-78, 5 s. ISBN 978-80-7043-806-0

Sojka, Petr.
Digitisation Workflow in the Czech Digital Mathematics Library.
Math-for-Industry, Kyushu, Japan : Faculty of Mathematics,Kyushu University, 2009, 22, od s. 272-280, 9 s. ISSN 1881-4042. 2009.

Sojka, Petr.
Languages of Mathematics Random Walking in the Mathematics of Languages.
In RASLAN 2009 Proceedings. prvni. Brno : Masaryk University, 2009. od s. 127-133, 7 s. ISBN 978-80-210-5048-8.

Číslo aktivity

26/09

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity
Rozšířená verze systému WebGen.
Zahájení aktivity
1.1.2009
Ukončení aktivity
31.12.2009
Popis aktivity
Došlo k testování a ověření použitých dialogových strategií na jak vidoucích, tak nevidomých uživatelích a připomínky byly zapracovány do systému WebGen. Byl navržen postup, jak umožnit snažší vytváření dialogových rozhraní přímo z popisovačů požadovaných dat. Výsledky byly publikovány na konferencích USAB09 (viz skutečné prostředky ověření) a Second International Conference on ICT & Accessibility. Dále byly navrženy postupy pro editaci stávajících prezentací vytvořených pomocí systému WebGen, které budou zaslány k prezentaci na konferenci ICCHP 2010. Dále je průběžně vytvářena anotovaná databáze grafických objektů.
Skutečné Indikátory dosažení - výsledky aktivity
Výsledkem aktivit je rozšířená verze systému WebGen s podporou přidávání nových prezentací dostupná na stránkách laboratoře (lsd.fi.muni.cz). Výstupy testů a výstupy testování systému WebGen a anotované databáze grafických objektů nevidomými uživateli byly prezentovány na konferencích USAB09 (viz skutečné prostředky ověření) a Second International Conference on ICT & Accessibility. Dále je na stránkách laboratoře (lsd.fi.muni.cz) dostupná nová verze systému WebGen.
Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity
Dosažená výsledky byly prezentovány na pracovních setkáních a formou publikací ve sbornících:

Bártek, L.: Generating Dialogues from the Description of Structured Data.
In HCI and Usability for e-Inclusion. Heidelberg : Springer-Verlag, 2009. od s. 227-235, 9 s. ISBN 978-3-642-10307-0

Kopeček, I. - Ošlejšek, R.: Accessibility of Graphics and E-learning. In Proceedings of the Second International Conference on ICT & Accessibility. Hammamet : Art Print, 2009. od s. 157-165, 9 s. ISBN 978-9973-37-516-2.

Číslo aktivity
27/09
Ke kterému dílčímu cíli se aktivita vztahuje
3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...
Název (cíl)aktivity
Anotátor grafických objektů
Zahájení aktivity
1.1.2009
Ukončení aktivity
31.12.2009
Popis aktivity
Detailnější rozpracování architektury anotátoru, vytvoření grafických ontologií grafické databáze s příhlédnutím k potřebám systému. Implementace první verze software pro anotaci grafických objektů.
Skutečné Indikátory dosažení - výsledky aktivity
Bylo provedeno rozpracování architektury anotátoru, vytvoření struktury hierarchických grafických ontologií.
Byla vytvořena základní anotovaná databáze grafických objektů a podpora procesu anotace včetně dialogového rozhraní pro práci s databází. Bylo rovněž implementováno propojení anotované databáze se systémem Webgen pro generování webovských presentací. Proběhlo testování procesu anotace objektů ve formátu SVG. Bylo implementováno zpřístupnění grafických objektů pomocí dotazovacího jazyka založeného na přirozeném jazyce.
Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity
Výsleky byly publikovány ve sbornících z mezinárodních konferencí. Byl vytvožen software pro podporu anotace

grafických objektů.
Publikace:
Bártek, L: Generating Dialogues from the Description of Structured Data. In Proceedings of the HCI and Usability for e-Inclusion Int. Conf., Springer Verlag, 2009, p. 227-235.

Kopeček, I., Ošlejšek, R., Plhák, R., Tiršel, F.: Detection and Annotation of Graphical Objects in Raster Images within the GATE Project. In Proceedings of the 2009 Int. Conf. on Internet Computing ICOMP 2009, CSREA Press, 2009, p. 285-290.

Číslo aktivity
28/09
Ke kterému dílčímu cíli se aktivita vztahuje
3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...
Název (cíl)aktivity
Integrace modelů vyhodnocování autoritativnosti v grafech
Zahájení aktivity
4.1.2009
Ukončení aktivity
30.6.2009
Popis aktivity
Aktivita vznikla dodatečně s potřebou syntetizovat procesy spojené s aktivitami 2007-09 (Návrh a implementace modelu webgrafu s cílem určit autoritativnost uzlů zohledňující skryté vazby komponent) a 2007-10 (Vyhodnocení výsledků navržených algoritmů pro analýzu struktury Webu) a shrnout jejich postupy a výsledky ucelenou formou. Součástí integračního úsilí byly rozsáhlé rešerše v dostupné literatuře a vypracování současného stavu znalostí v oblasti dolování z webu (web mining) a dolování z dat (data mining). Vzniklá metodologie zjišťování důležitých uzlů v orientovaných grafech nejrůznějšího charakteru je obecná a tedy vhodná k aplikaci v jakýchkoli prostředích, které můžeme modelovat jako orientovaný graf.
Skutečné Indikátory dosažení - výsledky aktivity
Výsledkem aktivity je obecná metodologie zjišťování důležitosti uzlů v prostředích modelovaných jako orientovaný graf, např. webový graf, citační graf autorů, publikací apod. Součástí je rovněž původní písemná zpráva o současném stavu oboru dolování informací z webu.
Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity
Výsledky aktivity byly prezentovány v monografii:
Fiala D. Web Mining Methods for the Detection of Authoritative Sources: Theory and Practice. VDM Verlag, Saarbrücken, Germany, 2009,
a v příspěvku na konferenci:
Jezek K.: Extended Formal Model for Ranking of Authoritative Resources. Proccedings XXII Mezdunarodnoj Naucnoj Konferencii MMTT-22, ISBN 978-5-91116-087-2 (Tom 7), pp.193-195, Pskov 2009.

Číslo aktivity
29/09
Ke kterému dílčímu cíli se aktivita vztahuje
3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...
Název (cíl)aktivity
Metody automatického posouzení kvality hlasu na základě prozodických charakteristik
Zahájení aktivity
15.1.2009
Ukončení aktivity

15.12.2009

Popis aktivity

Aktivita navazovala na výsledky roku 2008 a zabývala se automatickým posouzením kvality hlasu na základě prozodických charakteristik. Kvalita hlasu je rozhodujícím faktorem v porozumění spontánního dialogu. Integrace získaných výsledků do rozpoznávače přispěje ke zvýšení jeho přesnosti. Postup ocenění hlasových funkcí spočívá v subjektivním, ale hlavně objektivním posouzení jednotlivých složek hlasu. Základem je podrobné laryngostroboskopické vyšetření. Provádělo se měření hlasového pole a posouzení frekvenčního i dynamického rozsahu hlasu, včetně stanovení základní hlasové frekvence. Pomocí spektrální analýzy byla posuzována bohatost a kvalitu hlasu. Dále byla vytvořena experimentální databáze pro sledování poruch řeči.

Skutečné Indikátory dosažení - výsledky aktivity

Výsledkem aktivity jsou nové metody automatického posouzení kvality hlasu na základě prozodických charakteristik a návrh experimentálního systému pro sledování neurologických příčin poruch řeči.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledky této aktivity byly publikovány na dvou mezinárodních konferencích a v jejich sbornících (viz odst. 4.1.2).

2.2.2. AKTIVITY NEUSKUTEČNĚNÉ v roce 2009

Číslo aktivity
Ke kterému dílčímu cíli se aktivita vztahuje
Název (cíl)aktivity
Zahájení aktivity
Ukončení aktivity
Popis aktivity
Důvody, proč se aktivitu nepodařilo uskutečnit

2.3.NÁKLADY PROJEKTU - 2009

2.3.1. NÁKLADOVÉ TABULKY ZA JEDNOTLIVÉ SUBJEKTY

Rok	2009
Typ	skutečné
Organizace	Západočeská univerzita v Plzni
Role organizace	příjemce - koordinátor

POLOŽKA UZNANÝCH NÁKLADŮ tis. Kč	Náklady skutečně vynaložené tis. Kč	z toho skutečně hrazené z účelové podpory tis. Kč
F1. - Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	**2965**	**2945**
F2. - Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	**0**	**0**
F3. - Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	**0**	**0**
F4. - Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	**102**	**0**
F5. - Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	**23**	**0**
F6. - Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	**252**	**0**
F7. - Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	**303**	**50**
F8. - Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, enegii a služby neuvedené výše	**350**	**0**
F9. CELKEM	**3995**	**2995**

	PŘEVOD DO fondu tis. Kč	POUŽITÍ Z fondu tis. Kč
F0. - Zúčtování s Fondem účelově určených prostředků	**70**	**0**

	ZDROJE FINANCOVÁNÍ CELKEM tis. Kč	- z toho Účelová podpora (DOTACE) tis. Kč	- z toho Ostatní veřejné zdroje tis. Kč	- z toho Neveřejné zdroje tis. Kč
Z9.	**3995**	**2995**	**0**	**1000**

Rok	2009
Typ	skutečné
Organizace	Masarykova univerzita
Role organizace	příjemce

POLOŽKA UZNANÝCH NÁKLADŮ tis. Kč	Náklady skutečně vynaložené tis. Kč	z toho skutečně hrazené z účelové podpory tis. Kč
F1. - Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	**1918**	**1610**
F2. - Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	**0**	**0**
F3. - Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	**24**	**24**
F4. - Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	**8**	**6**
F5. - Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	**0**	**0**
F6. - Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	**0**	**0**
F7. - Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	**200**	**110**
F8. - Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, enegii a služby neuvedené výše	**200**	**0**
F9. CELKEM	**2350**	**1750**

	PŘEVOD DO fondu tis. Kč	POUŽITÍ Z fondu tis. Kč
F0. - Zúčtování s Fondem účelově určených prostředků	**0**	**0**

	ZDROJE FINANCOVÁNÍ CELKEM tis. Kč	- z toho Účelová podpora (DOTACE) tis. Kč	- z toho Ostatní veřejné zdroje tis. Kč	- z toho Neveřejné zdroje tis. Kč
Z9.	**2350**	**1750**	**0**	**600**

2.3.2. NÁKLADOVÁ TABULKA ZA PROJEKT

Rok	2009
Typ	skutečné
PROJEKT	2C06009 - CELKEM

POLOŽKA UZNANÝCH NÁKLADŮ tis. Kč	Náklady skutečně vynaložené tis. Kč	z toho skutečně hrazené z účelové podpory tis. Kč
F1. - Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	**4883**	**4555**
F2. - Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	**0**	**0**
F3. - Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	**24**	**24**
F4. - Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	**110**	**6**
F5. - Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	**23**	**0**
F6. - Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	**252**	**0**
F7. - Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	**503**	**160**
F8. - Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, enegii a služby neuvedené výše	**550**	**0**
F9. CELKEM	**6345**	**4745**

	PŘEVOD DO fondu tis. Kč	POUŽITÍ Z fondu tis. Kč
F0. - Zúčtování s Fondem účelově určených prostředků	**70**	**0**

	ZDROJE FINANCOVÁNÍ CELKEM tis. Kč	- z toho Účelová podpora (DOTACE) tis. Kč	- z toho Ostatní veřejné zdroje tis. Kč	- z toho Neveřejné zdroje tis. Kč
Z9.	**6345**	**4745**	**0**	**1600**

2.3.3. ZDŮVODNĚNÍ ZMĚN V ČERPÁNÍ

Dopisem z 15.10.2009 bylo MŠMT požádáno o povolení změny v položkovém členění uznaných nákladů. Povolená změna se týkala řešitelského pracoviště ZČU a neměnila celkově uznané náklady. Níže uvedená úprava byla schválena formou dodatku č.1/2009

Rok 2009:

-Snížení položky F7 „Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu“ o 150.000,-Kč (tj. na 300.000,-Kč)

-Navýšení položky F6 „Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu“ o 80.000,- Kč(tj. na 180.000,-Kč).

-Zbývajících 70.000,-Kč předpokládám převést do FÚUP (fond účelově určených prostředků) a jejich vyčerpání v roce 2010.

Rok 2010:

-Navýšení položky F6 „Náklady nebo výdaje na zveřejnění výsledků“ v roce 2010 o 70.000,-Kč (viz FÚUP), tj. na 170.000,-Kč.

-V roce 2010 snížení položky F7 „Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu“ o 150.000,-Kč (tj. na 350.000,-Kč) a zvýšení v roce 2010 o tuto částku položku F1 „Osobní náklady“ (tj. na 3115.000,-Kč).

Centrum ZPJ FI MU po dohodě s koordinátorem ZČU provedlo navýšení mzdových nákladů o částku 68 tisíc, přičemž stejnou částku byly sníženy provozní náklady projektu. Důvodem bylo navýšení tarifních platů v roce 2009 a zvýšení kvalifikace některých pracovníků. Provedený přesun financí představoval změnu o 2.8% z celkových nákladů na rok 2009 a navýšení mezd o 3.7%. Ze stejného důvodu plánujeme přesun částky 53 tisíc z provozních nákladů do mzdových prostředků i na rok 2010. Čerpaná částka na cestovné byla zvýšena o 20 tisíc, o tuto částku byly nižší provozní náklady. Celkové náklady projektu pro Masarykovu univerzitu byly čerpány v souladu se schváleným plánem.

2.3.4. NEVYUŽITÉ FINANČNÍ PROSTŘEDKY

Nevyužité prostředky položky F6 řešitelského pracoviště ZČU ve výši 70 tis. Kč, byly převedeny do fondu účelově určených prostředků a budou v souladu s povolenou úpravou (viz bod 2.3.3) použity k navýšení položky F6 v roce 2010.

2.3.5. Seznam hmotného a nehmotného majetku pořízeného za sledované období

3. ZÁMĚR A NÁVRHY PRO NÁSLEDUJÍCÍ OBDOBÍ - rok 2010

3.1. PROJEKTOVÝ TÝM A ŘEŠITELSKÉ TÝMY

3.1.1. PROJEKTOVÝ TÝM

IČ organizace	49777513
Obchodní jméno - název	**Západočeská univerzita v Plzni**
Zkratka názvu	ZČU
Role organizace	příjemce - koordinátor
Vazba na organizaci	00216224
Druh organizace	Veřejná nebo státní vysoká škola (zákon č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (o vysokých školách)
Adresa sídla, spojení na organizaci	
- ulice, čp./č.or.	Univerzitní 8/
- PSČ, obec	30614 Plzeň
- stát	Česká republika
- telefon	377 631 111
- http://	www.zcu.cz
Bankovní spojení	
-DIČ	CZ49777513
- banka kód, název	0100 - Komerční banka, a.s., Plzeň
- číslo účtu, sp.symbol	4811530257,
Statutární zástupce	
- titul před, jméno, příjmení, titul za	Doc. Ing. Josef Průša CSc.
- funkce	rektor
- telefon	377631000
- mobil	606665105
- fax	377631002
- email	rektor@rek.zcu.cz

IČ organizace	00216224
Obchodní jméno - název	**Masarykova univerzita**
Zkratka názvu	MU
Role organizace	spolupříjemce
Vazba na organizaci	49777513
Druh organizace	Veřejná nebo státní vysoká škola (zákon č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (o vysokých školách)
Adresa sídla, spojení na organizaci	
- ulice, čp./č.or.	Žerotínovo náměstí 617/ 9
- PSČ, obec	60177 Brno
- stát	Česká republika
- telefon	549 491 1111
- http://	www.muni.cz
Bankovní spojení	
-DIČ	CZ00216224
- banka kód, název	0100 - Komerční banka Brno-město
- číslo účtu, sp.symbol	85636621,
Statutární zástupce	
- titul před, jméno, příjmení, titul za	Prof. PhDr Petr Fiala PhD
- funkce	rektor
- telefon	549491001
- mobil	
- fax	
- email	rektor@muni.cz

3.1.2. ŘEŠITELSKÝ TÝM

Celé jméno, RČ	**Bártek Luděk Mgr.** 7201083791 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 3215 bar@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita Fakulta informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	

Celé jméno, RČ	**Brada Přemysl Ing. PhD. MSc.** 7007012111 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	3772435 brada@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

Celé jméno, RČ	**Dostal Martin ing.** 8409092054 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632452 604796109 madostal@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	25

Celé jméno, RČ	**Ekštein Kamil Ing. PhD.** 7705302011 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 kekstein@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Fiala Dalibor Ing PhD.** 8003235845 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632429 dalfia@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	25

Celé jméno, RČ	**Habernal Ivan Ing.** 8307051764 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 377 632 402 habernal@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	50

Celé jméno, RČ	**Hejtmánek Jan Ing.** 8211012095 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 377 632 402 hejtman2@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	12.5

Celé jméno, RČ	**Horák Aleš RNDr. Ph.D.** 7409014250 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 4377 hales@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita Fakulta informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	

Celé jméno, RČ	**Ježek Karel doc. Ing. CSc.** 420617110 CZ
Role osoby při řešení projektu	řešitel
Spojení	377 632 475 jezek_ka@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Ježek Petr Ing.** 8302082734 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632401 723123877 jezekp@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni FAV - KIV
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	12.5

Celé jméno, RČ	**Klečková Jana doc. Dr. Ing.** 496108095 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 421 kleckova@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

Celé jméno, RČ	**Konopík Miloslav Ing. PhD.** 8103261782 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 konopik@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	100

Celé jméno, RČ	**Kopeček Ivan doc. RNDr. CSc.** 490303075 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 3861 kopecek@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita Fakulta informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	

Celé jméno, RČ	**Král Pavel Ing. PhD.** 7603172049 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632454 pkral@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Krčmář Lubomír Ing.** 8408221228 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632452 lkrcmar@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	25

Celé jméno, RČ	**Krutišová Jana Ing.** 5955160046 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 413 krutisova@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

Celé jméno, RČ	**Matoušek Václav prof. Ing. CSc.** 480613108 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 471 matousek@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Mautner Pavel Ing. PhD.** 6505222592 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 441 mautner@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Mouček Roman Ing. PhD.** 7607072000 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 441 moucek@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Pala Karel doc. PhDr. CSc.** 390615416 CZ
Role osoby při řešení projektu	spoluřešitel
Spojení	549 49 5616 pala@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita Fakulta informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	

Celé jméno, RČ	**Pomikálek Jan Mgr.** 7910090419 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 1864 xpomikal@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita Fakulta informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	

Celé jméno, RČ	**Ptáčková Helena** 7059142079 CZ
Role osoby při řešení projektu	osoba autorizovaná k finančním záležitostem
Spojení	377 632 463 377 632 402 ptackova@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	5

Celé jméno, RČ	**Rambousek Adam Bc.** 8110225233 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	xrambous@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita Fakulta informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	

Celé jméno, RČ	**Rychlý Pavel Mgr. Ph.D.** 7301235359 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 6399 pary@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita Fakulta informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	

Celé jméno, RČ	**Sojka Petr doc. RNDr. Ph.D.** 6309171000 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549496966 sojka@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita Fakulta informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	

Celé jméno, RČ	**Zíma Martin Ing. PhD.** 7405042073 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632431 zima@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

3.1.3. ZMĚNY V PROJEKTOVÉM A ŘEŠITELSKÝCH TÝMECH - rok 2010

Pč.	Typ	Popis
*		

3.2. ČASOVÝ POSTUP PRACÍ - rok 2010

3.2.0. PŘEHLED DÍLČÍCH CÍLŮ PLÁNOVANÉ 2010

	Číslo	Dílčí cíl podrobně	Datum plnění
	1	**Dílčí cíl** Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřování algoritmů komunikace s www prostředím. **Indikátory dosažení - výsledky dílčího cíle** a) Vytvoření uživatelského rozhraní pro hlasový vstup / příp. výstup, které bude použito pro komunikaci se sémantickým webem, a pro jeho podporu vytvoření robustního ASR systému pro inflexní jazyky. K tomu bude nutno vytvořit kvalitní korpus pro ASR a z něj extrahovat dostatečné množství trénovacích dat. V jednotlivých etapách bude v průběhu let 2006 – 2007 vytvořen: - kvalitní audio-korpus pro natrénování systému ASR, - korpus pro natrénování jazykových modelů. b) Příprava datových kolekcí a pomocných rutin vyhledávacího systému ve vícejazyčných korpusech, včetně prostředků pro zpřesňování uživatelských dotazů pomocí thesauru a nástrojů pro disambiguaci víceznačných slov, na bázi klient/server aplikace. Jednotlivé dílčí výsledky řešení projektu lze charakterizovat takto: - vytvoření multijazykových korpusů – základní výběr zahrnuje angličtinu a češtinu, dle možností alespoň některé úlohy plánujeme provádět i se slovenštinou (zajímavá je blízkostí k češtině) a němčinou, - metoda automatického rozpoznání jazyka – kombinace „stop slov“ a frekvenčních znakových metod. c) Příprava datových kolekcí a modulů pro filtraci a sumarizaci textů: - vytvoření sumarizačních korpusů (pro angličtinu plánujeme využít standardních korpusů, např. DUC a pro češtinu bude vytvořen vlastní, složený vesměs z textů novinových článků, - sumarizace textů založená na latentní sémantické analýze (LSA), vytvoření anotované kolekce pro sumarizátor založený na LSA - vytvoření vícejazyčných korpusů, - rozšíření standardních textových korpusů o korpusy závadných dokumentů pokrývající problematická témata definovaná v zadání. d) Korpus syntaktických stromů (treebank): - korpus bude morfologicky označkován a zjednoznačněn, - bude v něm vyznačena závislostní struktura věty i jednotlivé větné složky včetně koreferenčními vztahy, - korpus bude z části založen na existujícím PDT. e) Korpus vzorových přepisů vybraných vět a jejich sémantické reprezentace: - text korpusu bude podmnožinou korpusu syntaktických stromů, - ve stromech budou vyznačeny významy z dostupných ontologií (WordNet), - věty budou rozšířeny o logické formy. f) Doplnění morfologického značkovače o robustní hádací proceduru, která bude spolehlivě přiřazovat morfologické značky i neznámým slovům. **Prostředky ověření - Forma zpracování a předání výsledku dílčího cíle** Jedná se o vytvoření podpůrného aparátu, bez něhož nelze další zamýšlené cíle projektu dosáhnout. Vytvořeny budou proto korpusy v podobě rozsáhlých datových souborů se specifickou strukturou a organizací a pro jejich údržbu a prohledávání budou vyvinuty speciální softwarové nástroje. Výsledky budou soustředěny do soustavy datových souborů a její obsah prezentován formou publikace na konferencích a v průběžných výzkumných zprávách. **Kritické poedpoklady dosažení dílčího cíle** Rizikové faktory ovlivňující náplň dílčího cíle „1“ a nástin jejich řešení jsou následující: RF1: Během zpracování korpusů a korpusových nástrojů se vyskytnou další korpusy obsahující srovnatelná data. Řešení: Korpusy pro český jazyk vznikají v ČR na celkem pěti pracovištích, která udržují těsné kontakty a výsledky výzkumu si vzájemně vyměňují nebo se o nich poměrně obsáhle informují. Navíc je třeba rozlišovat mezi korpusy psanými (textovými) a řečovými. Řečové korpusy vznikají prakticky jen na pracovištích v Plzni, Brně a Liberci, z nichž dvě se na řešení tohoto projektu budou podílet. Navíc vznik jakéhokoli dalšího korpusu je pozitivním jevem, neboť v tomto oboru	- 31.12.2007

		více než kdekoli jinde platí, že vhodných dat není nikdy dostatek. Tudíž korpusy vytvořené v rámci navrhovaného projektu budou v každém případě využity i dalšími pracovišti. V případě cizojazyčných korpusů budou využívány korpusy, které jsou k dispozici v systému ELRA (European Language Resources Association). RF2: Nepodaří se získat dostatek materiálů, resp. mluvčích, pro vytvoření textových, resp. audiokorpusů. Řešení: Tento rizikový faktor nebude mít zřejmě přílišnou váhu, neboť již současný web poskytuje doslova nepřeberné množství textového materiálu, z nichž lze za použití vhodných vyhledávacích metod vybrat dostatečné množství materiálu pro vytvoření korpusu. V případě řečových korpusů nejde ani tak o problém nalezení vhodné množiny dat nebo množiny vhodných mluvčích, nýbrž kritickým faktorem je čas. Pořizování řečových dat a zejména jejich následné zpracování (třídění, anotace, apod.) vyžaduje značné množství času, avšak riziko lze úspěšně odstranit kvalitním managementem projektu. RF3: V průběhu naplňování dílčího cíle projektu se vyskytne komerční software řešící problematiku pořizování korpusů. Řešení: Pokud se nějaký software vyskytne a bude využitelný, nebude díky modularitě předpokládaného programového vybavení příliš obtížné ho do vytvářeného software začlenit. Pravděpodobnost jeho výskytu v dohledné době je však minimální.	
	2	**Dílčí cíl** Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka. **Indikátory dosažení - výsledky dílčího cíle** a) Návrh formalismu pro popis sémantiky na rozsáhlejší doméně, návrh vhodně strukturovaného sémantického popisu dotazů uživatelů, eventuálně vytvoření vlastního hierarchického systému relací mezi lexémy pro zaručení generalizační schopnosti systému. b) Vytvoření ontologií pro aplikaci formalismu popisujícího sémantiku. Jednotlivými výsledky budou: - návrh ontologie, sémantických konceptů – datový formát XML, vytvoření UML modelu, - návrh ohodnocení jednotlivých konceptů vektorem sémantických příznaků, a to jak doménových, tak obecnějšího charakteru, - návrh soustavy vektorů ohodnocení jednotlivých konceptů. c) Vytvoření multilingválního sumarizačního systému včetně rezoluce anafor a komprese souvětí, jeho zakomponování do prostředí pro vyhledávání a vývoj metod ohodnocování jeho kvality, návrh metod disambiguace v multijazykovém prostředí s využitím kontextu, thesauru a pravděpodobnostních metod: - sumarizační systém obohacený o kompresi souvětí, - systém rezoluce anafor a jeho využití při sumarizaci – pro angličtinu bude využit systém GuiTAR, vytvořený na univerzitě Essex (Anglie), pro češtinu bude na základě poznatků získaných na českých pracovištích vytvořen vlastní systém, - metoda hodnocení kvality sumarizátorů na základě LSA. d) Vývoj nových, dokonalejších modelů elektronických dokumentů tak, aby při použití textových klasifikačních algoritmů bylo dosaženo co nejlepších výsledků při rozpoznávání tématu, rozpoznávání spamových emailů, detekci dokumentů se závadným obsahem apod. e) Vytvoření metodologie a nástrojů pro analýzu webových dokumentů. **Prostředky ověření - Forma zpracování a předání výsledku dílčího cíle** Při naplňování tohoto dílčího cíle půjde o vytvoření základního teoretického podpůrného aparátu, bez něhož nebude možné další kroky realizovat. Jediný tento dílčí cíl bude mít charakter spíše základního výzkumu – půjde o vývoj metod, metodologií a formálních modelů pro návrh zamýšleného komunikačního rozhraní, avšak součástí výzkumných prací bude též experimentální implementace a vytvoření softwarových nástrojů pro evaluaci vyvíjených metod a formalismů. Výsledky budou shrnuty do písemných dokumentů a prezentovány téměř výhradně formou publikací na konferencích, v odborných časopisech a v průběžných výzkumných zprávách. **Kritické poedpoklady dosažení dílčího cíle** Rizikové faktory ovlivňující dosažení dílčího cíle „2“ a nástin jejich řešení mohou být následující: RF1: Nepotvrzení či neplatnost výzkumných hypotéz poskytujících základ pro vytvoření formalismů a modelů. Řešení: Plánovaný dílčí cíl zde nestojí na jediné výzkumné hypotéze, nýbrž na teoretickém základu	- 31.12.2008

		návrhu komunikačních systémů. Využito bude jak dosavadních poznatků z návrhu existujících komunikačních rozhraní a systémů pro interakci člověka s počítačem, tak i poznatků z psychologie komunikace a doporučení TC.13 IFIP (for HCI). Základním rizikem proto bude opět časový faktor, který lze výrazně omezit dobrým managementem projektu. RF2: Nedostatečná erudice členů týmu pro vývoj formálních prostředků. Řešení: Tento rizikový faktor nebude mít zřejmě přílišnou váhu, neboť oba participující týmy jsou složeny minimálně z poloviny ze starších zkušených výzkumníků, z nichž někteří se předmětnou oblastí zabývají 25 i více let, z druhé části pak z mladých perspektivních pracovníků, kteří buď vyrostli anebo se podíleli na řešení podobné problematiky a potřebné teoretické základy oboru již získali, zejména v doktorandském studiu.	
	3	**Dílčí cíl** Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. **Indikátory dosažení - výsledky dílčího cíle** a) Implementace uživatelského rozhraní pro hlasovou komunikaci se sémantickým webem –součástí výsledku budou: - implementace LVCSR rozpoznávače, - natrénování akustických a jazykových modelů, - implementace nahrávacího modulu se stochastickým modelem detekce řečového signálu, - implemetace parametrizátoru na bázi MFCC, - návrh a implementace modulu pro akustické modelování založeného na umělých neuronových sítích nebo směsích Gaussových funkcí, - návrh a implementace efektivního dekódovacího algoritmu, který dokáže pracovat s gramatikami a stochastickými jazykovými modely, - programová realizace a ověření funkčních vlastností robustního ASR systému pro inflexní jazyky. b) Systém pro extrakci významu ze spontánních promluv – dílčími kroky k dosažení tohoto dílčího cíle budou: - návrh a realizace optimální řečové databáze, - návrh systému sémantického značkování řečových dat, - báze znalostí umožňující automatizované či automatické značkování spontánních promluv uložených v databázi, - implementace stochastických sémantických gramatik pro automatickou sémantickou analýzu dotazu uživatele, - využití hierarchické ontologie pro tvorbu strukturalizovaného popisu dotazů uživatele a pro zajištění schopnosti zobecňování z natrénovaných dat, - aplikace metod mělkého (shallow) parsingu promluv pro částečnou analýzu dotazů uživatele. c) Vytvoření komfortního uživatelského rozhraní pro práci se sémantickým webem – součástí tohoto dílčího cíle bude: - návrh příslušného dialogového manageru akceptujícího tzv. kombinovanou iniciativu ve vedení dialogu (mixed initiative), - vytvoření robustního systému pro efektivní a časově nenáročné vyhledávání dat v řečové databázi, - vytvoření robustního a spolehlivého modelu sémantické hierarchie a jeho implementace. d) Aplikace a modifikace OWL standardu v českém prostředí. e) Aplikace klasifikačních metod v multijazykovém prostředí. f) Kompletace multilingválního sumarizačního systému včetně rezoluce anafor a komprese souvětí. g) Algoritmy vhodné pro generování itemsetů a n-gramů a ověření jejich úspěšnosti pro klasifikaci textových dokumentů. h) Výchozí algoritmy pro vyvozování nových znalostí z informací získaných z volného textu. i) Prototyp programu pro přiřazování logických formulí větám z volného textu. **Prostředky ověření - Forma zpracování a předání výsledku dílčího cíle** V dílčím cíli „3“ jde o vytvoření souboru programových produktů, které vzniknou implementací teoretických metod a formalismů vytvořených v rámci dílčího cíle „2“. Výsledky budou mít jednoznačně aplikační charakter, i když vesměs půjde jen o experimentální software, bez něhož nelze metody a modely verifikovat. Výsledky však bude možno předat i dalším zájemcům, protože se předpokládá úplná dokumentace vytvořeného programového vybavení. Výsledky budou prezentovány jako balíky experimentálního software a metod, dále budou publikovány na konferencích, v průběžných výzkumných zprávách, případně také zveřejněny formou speciálních	- 31.12.2009

		letáků, v tisku a uvažuje se též o možnosti předvedení na specializovaných veletrzích a výstavách. **Kritické poedpoklady dosažení dílčího cíle** Rizikové faktory ovlivňující dosažení dílčího cíle „3“ a nástin jejich řešení: RF1: V průběhu projektu přestane být o vytvářené přístupové technologie zájem a pracoviště účastníci se na řešení projektu se tak ocitnou bez reálné využitelnosti svých výsledků. Řešení: Současným trendem je naopak příklon k využívání multimediálních a multimodálních dat, ukládání velkých množství dat a informací na běžných počítačových prostředcích, sílí propojování informačních technologií s rozhlasovým a televizním vysíláním, streamovanými médii a mobilními komunikacemi. Nové hardwarové prostředky budou vyžadovat nové technologie přístupu k datům, přičemž preferována bude komunikace v přirozeném jazyce, ať už psanou nebo mluvenou formou. Vyvíjené programové prostředky tento trend jednoznačně podpoří a proto je toto riziko za dobu řešení projektu téměř nulové. RF2: V průběhu řešení projektu se vyskytne komerční software řešící problematiku srovnatelnou s předpokládanými výsledky projektu. Řešení: Komerční řešení využívající přístup k datům na webu prostřednictvím přirozeného jazyka jsou dosud v plenkách a komerční sféra naopak aktivně vyhledává zajímavé práce z akademické sféry. Proto je toto riziko minimální, očekáváme naopak velký zájem z komerční sféry. RF3: Časové faktory ovlivňující zpracování software. Řešení: Při implementaci a programové realizaci metod vyvinutých v rámci dílčího cíle „2“ může dojít k určité časové tísni vlivem nevhodně zvolených implementačních nástrojů, eventuálně nezkušeností některých mladších členů týmu. Riziko je však minimální, neboť řešitelský kolektiv je složen vesměs ze zkušených výzkumníků a mladých pracovníků, kteří již obdobné, i když jednodušší systémy v minulosti vytvářeli a implementovali. Časový faktor lze navíc výrazně ovlivnit dobrým managementem projektu.	
	4	**Dílčí cíl** Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí. **Indikátory dosažení - výsledky dílčího cíle** a) Testování a ověřovací provoz implementovaného hlasového rozhraní – součástí bude - otestování zpracovaného LVCSR rozpoznávacího systému, - ověření funkčních vlastností robustního ASR systému na vhodné množině uživatelů, - otestování vyvinutých metod automatické sémantické analýzy dotazů. b) Ověření funkčních vlastností vytvořených ontologií a hierarchického systému relací mezi lexémy pro zaručení generalizační schopnosti systému analýzy sémantiky, c) Ověření vlastností algoritmů pro klasifikaci a analýzu dat na různých typech dokumentů. d) Otestování a ověření navržených metod na konkrétních typových řešeních, např. na přístupu k webovým stránkám výzkumných a vzdělávacích institucí. e) Vyhodnocení úspěšnosti jednotlivých fází analýzy volného textu od morfologické úrovně až po převod do logických formulí. **Prostředky ověření - Forma zpracování a předání výsledku dílčího cíle** Náplní dílčího cíle „3“ je provedení rozsáhlých testů (tzv. field experiments) vyvinutých metod, metodologií, modelů a vytvořeného souboru programových produktů. Předpokládá se testování produktů na obvyklých třech skupinách uživatelů – v prvním kroku budou vlastnosti systémů a metod prověřovány úzkou skupinkou řešitelů projektu, ve druhém kroku bude testovací množina uživatelů vytvořena ze spolupracovníků, kteří však s řešením projektu neměli nic společného a o výsledcích řešení jsou jen velmi kuse informováni, a teprve ve třetím kroku bude systém testován libovolnými uživateli, tzv. „lidmi z ulice“. Zčásti však v tomto kroku budou využiti studenti, kteří všeobecně mají tendenci takové systémy „pokořit“. Výsledky budou kompletně dokumentovány a z vyhodnocení experimentů budou vyvozovány příslušné závěry, tj. systém a jeho části budou průběžně doplňovány, upravovány a opětovně testovány. V závěru budou výsledky testování a ověřovacího provozu publikovány v časopisech, na konferencích a obšírně v závěrečné výzkumné zprávě. **Kritické poedpoklady dosažení dílčího cíle** Rizikové faktory ovlivňující dosažení dílčího cíle „4“ a možná řešení: RF1: V průběhu testů se projeví nedostatky v koncepci systému vedoucí k závažným problémům ve funkci systému.	- 31.12.2010

		Řešení: Řešitelský tým je složen z odborníků, kteří obdobné, i když jednodušší, systémy již vytvořili a mají z jejich tvorby nezanedbatelné zkušenosti. Tým byl dále doplněn o mladé pracovníky, kteří se podíleli na tvorbě řady produktů pro prezentace na webových stránkách a je jim problematika přístupu k webu velmi blízká. Riziko volby nevhodné koncepce je proto minimální. RF2: V průběhu testů se projeví nedostatky v implementaci systému a metod. Řešení: Obdobné jako předchozí rizikový faktor – řešitelský tým je složen z odborníků, kteří obdobné, systémy již vytvořili a mají i z jejich implementace poměrně rozsáhlé zkušenosti. Riziko závažných implementačních chyb je proto minimální, drobné nedostatky v implementaci bývají zpravidla v krátké době snadno odstranitelné. RF3: Nepodaří se vytvořit dostatečně reprezentativní množiny testovacích osob. Řešení: Ve vztahu k odstavci 3.3.3. (tři úrovně testování) je riziko nedostatečného vytvoření skupin testujících osob nepatrné – obě participující pracoviště jsou poměrně rozsáhlá a množinu osob testujících vlastnosti systému nebude problém vytvořit; ostatně bylo již ověřeno v minulosti na jednodušších úlohách. Otázka volby třetí skupiny osob je spíše otázkou vytvořeného přístupu k systému – zde se nabízejí dvě možnosti: Buď si osoby vhodné k testování systému vybírat podle určitých hledisek (bylo tak někdy postupováno v minulosti a osoby byly k testování zvány na řešitelské pracoviště) nebo zveřejnit přístupový portál systému a dovolit testování systému široké veřejnosti prostřednictvím internetu, popř. přes telefon (telefonní přístup je však v současných podmínkách omezen kvalitou spojení v mobilních sítích, resp. kvalita spojení je dána úrovní signálu v místech, kde se potenciální uživatel právě nachází, a výsledky testů jím mohou být zkresleny). Rizikový faktor může být opět minimalizován vhodnými rozhodnutími, resp. dobrým managementem projektu.	

3.2.1. AKTIVITY PLÁNOVANÉ NA DALŠÍ OBDOBÍ - rok 2010

Číslo aktivity

01/10

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

Nasazení systémů pro odpovídání otázek v přirozeném jazyce do reálného prostředí

Zahájení aktivity

4.1.2010

Ukončení aktivity

22.12.2010

Popis aktivity

Cílem této aktivity bude v návaznosti na aktivitu 2009-05 nasazení experimentálních systémů pro odpovídání otázek do reálného prostředí. Aktivita se soustředí na další postupné doplňování databází a odladění chyb systémů.

Plánované indikátory dosažení - očekávané výsledky aktivity

Výsledkem aktivity budou dva systémy pro odpovídání otázek nasazené na veřejně přístupném serveru v prostředí Internetu.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Systémy budou k dispozici v síti Internet. Data a zdrojové kódy budou uloženy v zálohovaném repository, dosažené výsledky budou postupně publikovány v odborných textech a na konferencích.

Číslo aktivity

02/10

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

Vývoj nových metod pro automatickou sémantickou analýzu dotazů

Zahájení aktivity

4.1.2010

Ukončení aktivity

22.12.2010

Popis aktivity

Aktivita úzce navazuje na aktivitu 2009-04. Jejím cílem je vylepšení stávajícího algoritmu sémantické analýzy a zohlednění bohatě inflexních jazyků (např. češtiny) použitím pokročilých metod předzpracování. Mezi navrhovanými metodami figuruje např. úprava metody morfologické analýzy. Taktéž bude brána v potaz chybovost vstupu. Pro zvýšení úspěšnosti identifikace lexikálních tříd bude využita vlastní databáze geografických a vlastních názvů.

Plánované indikátory dosažení - očekávané výsledky aktivity

Zlepšení úspěšnosti automatické sémantické analýzy výsledky budou porovnávány s ručně anotovanými daty.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledky vývoje algoritmu budou v závěru roku publikovány v odborné literatuře.

Číslo aktivity

03/10

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

Využití jazykových modelů v programu JLaser

Zahájení aktivity

4.1.2010

Ukončení aktivity

22.12.2010

Popis aktivity

V programu recognizeru JLaser je zatím použita pouze základní podpora pro jazykové modely. Aby systém rozpoznávání řeči mohl být nasazen do běžné praxe, je třeba do systému doplnit plnou podporu a otestovat úspěšnost nového ASR s jazykovými modely vůči staré verzi. Je obecně známo, že jazykové modely výrazně zvyšují úspěšnost rozpoznávání, což se (doufáme) podaří prokázat i pro rozpoznávání českých promluv.

Plánované indikátory dosažení - očekávané výsledky aktivity

Natrénované jazykové modely a implementace softwarového modulu pro recognizer JLaser.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Testy úspěšnosti rozpoznávání, ověření kvality natrénovaných jazykových modelů.

Číslo aktivity

04/10

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

Trénování akustických modelů

Zahájení aktivity

4.1.2010

Ukončení aktivity

30.9.2010

Popis aktivity

Z předešlého výzkumu vyplynula potřeba lepších svazovacích metod (clustering) pro slabikové akustické modely v systému rozpoznávání řeči. Techniky, které byly používány na triphonové a monophonové akustické modely fungují na slabikové úrovni pouze částečně. V dalším výzkumu tak budou dále vylepšeny techniky decision tree a data-driven clusteringu pro tvorbu akustických modelů.

Plánované indikátory dosažení - očekávané výsledky aktivity

Natrénované akustické a jazykové modely a jejich využití pro rozpoznávání spojitých českých promluv.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Testy úspěšnosti rozpoznávání spojitých českých promluv. Výsledky vývoje algoritmu budou postupně publikovány v odborných článcích.

Číslo aktivity

05/10

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

Kategorizace dokumentů neuronovou sítí

Zahájení aktivity

4.1.2010

Ukončení aktivity

22.12.2010

Popis aktivity

V počáteční fázi aktivity budou dokončeny testy s neuronovou sítí typu vícevrstvý perceptron (MLP), který byl navržen jako výstupní vrstva neuronového systému pro kategorizaci dokumentů. V další fázi se pak pozornost zaměří na vytvoření hierarchické struktury neuronových sítí SOM, které ve výsledném systému nahradí mapu slovních kategorií. Předpokládá se, že tato hierarchická struktura provede přesnější kategorizaci slov, což by mělo mít v

důsledku vliv i na výslednou kategorizaci dokumentů. Navržená síť bude natrénována jednak korpusy vytvořenými v rámci projektu, korpusy vlastními, vytvořenými např. z webových článků patřících do specifické domény (sport, počasí apod.). Navržená a implementovaná hierarchická struktura bude začleněna do neuronového systému pro kategorizaci dokumentů a výsledky kategorizace budou porovnány s výsledky dosaženými v předchozích aktivitách.

Plánované indikátory dosažení - očekávané výsledky aktivity

Výsledkem bude implementace hierarchické neuronové sítě SOM a její začlenění do stávajícího systému pro kategorizaci dokumentů.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Testy úspěšnosti kategorizace. Výsledky budou prezentovány v odborné literatuře, software bude veřejně dostupný v příslušném adresáři - http://www.kiv.zcu.cz/vyzkum/software/

Číslo aktivity

06/10

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

Sémantický web pro EEG/ERP doménu

Zahájení aktivity

4.1.2010

Ukončení aktivity

22.12.2010

Popis aktivity

Praktická realizace sémantického webu používající standardy a technologie RDF a OWL (pro realizaci bude využita již exisrtující databáze EEG/ERP experimentů). Ověření úspěšnosti transformace mezi relační databází (nebo objektovým modelem) a vyjadřovacími prostředky sémantického webu. Úprava uživatelského rozhraní a jeho rozšíření o možnosti zadání dotazu v přirozeném jazyce.

Plánované indikátory dosažení - očekávané výsledky aktivity

Aktualizace ontologie EEG/ERP domény, ověření a úpravy transformačního mechanismu pro výměnu metadat s relační databází (objektovým modelem) – statistika úspěšnosti dotazů do databáze EEG/ERP experimentů, rozšíření uživatelského rozhraní o možnost zadání dotazu v přirozeném jazyce.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledky budou prezentovány v odborné literatuře, software bude veřejně dostupný na internetových stránkách - http://www.kiv.zcu.cz/vyzkum/software/.

Číslo aktivity

07/10

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

Zhodnocení metod automatického rozpoznávání dialogových aktů

Zahájení aktivity

4.1.2010

Ukončení aktivity

15.12.2010

Popis aktivity

Aktivita, navazující na aktivitu 2009-07, se bude zabývat ověřením a zhodnocením navržených a implementovaných metod automatického rozpoznávání dialogových aktů. Funkce metod byla zatím ověřena na pouze dvou malých korpusech ve dvou evropských jazycích. Český korpus čítal celkem 2173 dialogových aktů, francouzský celkem 1518 dialogových aktů. Cílem aktivity proto bude ověření funkce metod na rozsáhlém korpusu, který bude čítat několik

desítek tisíc vět. Dále bude zapotřebí zjistit, zda by nebylo možné modifikované metody automatického rozpoznávání dialogových aktů použít k automatickému doplnění větné interpunkce. Tato informace není součástí výstupu recognizeru, ale mohla by napomoci při určování významu věty.

Plánované indikátory dosažení - očekávané výsledky aktivity

Natrénované robustní modely dialogových aktů.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Testy úspěšnosti, srovnání a diskuze výsledků rozpoznávání dialogových aktů jednotlivých metod, publikace dosažených výsledků na konferencích a v odborných časopisech.

Číslo aktivity

08/10

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Metody automatického rozpoznávání dialogových aktů na základě dialogové historie

Zahájení aktivity

15.1.2010

Ukončení aktivity

15.12.2010

Popis aktivity

V rámci předchozích aktivit byly využity k rozpoznávání dialogových aktů tyto hlavní zdroje informací: lexikální informace, prozódie a výraz obličeje. Tyto informace budou v dalším doplněny o tzv. dialogovou historii (časovou sekvenci po sobě jdoucích dialogových aktů). Cílem této aktivity bude zaměření na tuto v našich metodách zatím chybějících informaci, ověření dostupných metod, které tuto informaci využívají, a výběr nejvhodnější metody. Dále bude nutno rozšířit kolekci dosud zkoumaných paralingvistických aspektů promluvy s cílem vymezení dialogového aktu a rozpoznání sémantiky promluvy.

Plánované indikátory dosažení - očekávané výsledky aktivity

Implementace metod automatického rozpoznávání dialogových aktů využívajících dialogovou historii. Stanovení nové kolekce prozodických parametrů na základě datového skladu.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Experimentální ověření zvolených metod na českém korpusu dialogových aktů.

Číslo aktivity

09/10

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

Analýza citačních sítí autorů a institucí získaných z webových dat

Zahájení aktivity

4.1.2010

Ukončení aktivity

31.3.2010

Popis aktivity

Cílem této činnosti je navázat na aktivitu 2007-09 (Návrh a implementace modelu webgrafu s cílem určit autoritativnost uzlů zohledňující skryté vazby komponent) a aplikovat vyvinuté metody na data z databáze CiteSeer (http://citeseer.ist.psu.edu). Na rozdíl od dat z digitální knihovny DBLP, jež jsou sestavována ručně, jsou v počítačově utvářeném CiteSeeru některé informace navíc, např. adresy a instituce publikujících autorů. Také citací mezi publikacemi je nepoměrně více, a proto je zpracovávaná citační síť mnohem rozsáhlejší. To bude mít za následek práci s mnohem větším objemem dat, ale i bohatší výstupy výzkumu. Výsledkem činnosti budou tedy seznamy

autoritativních autorů, institucí a zemí a také porovnání seznamu autorů s autory vzešlými z aktivity 2007-09 při práci s DBLP.

Plánované indikátory dosažení - očekávané výsledky aktivity

Indikátorem dosažení výsledku je vytvoření citačních grafů autorů, institucí a zemí a to ve formě relační databáze. Na základě dotazů a výpočtů nad touto databází vzniknou seznamy (žebříčky) nejpopulárnějších entit.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledky experimentů budou publikovány formou výzkumných článků. Programové nástroje a relační databáze s daty budou uloženy v datovém úložišti projektu. Úspěšnost metod bude možno ověřit porovnáním s výsledky aktivity 2007-09 a s výsledky již publikovanými v odborné literatuře.

Číslo aktivity

10/10

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

Zahrnutí časového faktoru do citačních sítí

Zahájení aktivity

1.4.2010

Ukončení aktivity

22.12.2010

Popis aktivity

Záměrem této aktivity je zahrnout do metod vyvinutých v rámci aktivity 2007-09 (Návrh a implementace modelu webgrafu s cílem určit autoritativnost uzlů zohledňující skryté vazby komponent) faktor času a v závislosti na něm rozdílně ohodnocovat hrany v analyzovaných orientovaných grafech. Je zřejmé, že z takovéto „časové“ analýzy je třeba vyjmout webový graf, neboť odkazy na webových stránkách samy o sobě nenesou datum ani čas. Analýze tedy budou podrobeny pouze citační grafy autorů a jejich publikací, u nichž se zpravidla datum (alespoň rok) dají snadno zjistit. Použijí se při tom informace z digitálních knihoven DBLP (http://dblp.uni-trier.de/) a CiteSeer (http://citeseer.ist.psu.edu) získané v jiných aktivitách tohoto projektu. Základní myšlenka pro zahrnutí časového faktoru do výpočtu důležitosti uzlů grafu spočívá v tom, že citace následující po společné publikaci dvou citujících se autorů má menší váhu než citace předcházející společné publikaci.

Plánované indikátory dosažení - očekávané výsledky aktivity

Indikátorem dosažení výsledku jsou vytvořené žebříčky nejuznávanějších autorů, které reflektují jejich vliv na vědeckou komunitu „férověji“ než pouhé počty citací vlastních publikací.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledky experimentů budou publikovány formou výzkumných článků. Rovněž programové nástroje a relační databáze s daty budou uloženy v datovém úložišti projektu. Úspěšnost metod bude možno ověřit porovnáním s výsledky aktivity 2007-09 a s výsledky již publikovanými v odborné literatuře.

Číslo aktivity

11/10

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Extrakce informací z textů e-mailů typu Call for papers (žádost o zaslání příspěvků na konferenci)

Zahájení aktivity

4.1.2010

Ukončení aktivity

22.12.2010

Popis aktivity

Cílem plánované aktivity bude vytvoření prototypu programu pro extrakci informací z volného textu. Informace z textů e-mailů budou extrahovány z dat uložených ve formě xml. Extrahovaná data budou dále ukládána do databáze MySql a následně přehledně zobrazena na Webu.

Plánované indikátory dosažení - očekávané výsledky aktivity

Prototyp nástroje pro extrakci informací z textů e-mailů typu Call for papers.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Testování schopnosti extrakce informací bude možné ověřit prostřednictvím webového formuláře.

Číslo aktivity

12/10

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

Ověření a testování funkčnosti transformace XML dokumentu do logického programu

Zahájení aktivity

4.1.2010

Ukončení aktivity

22.12.2010

Popis aktivity

Bude vytvořena sada programů, které budou implementovat postupy navržené v aktivitě 2009-16. Tyto programy ověří, zda vstupní XML soubor vyhovuje podmínkám transformace na logický program a provedou předzpracování vstupního dokumentu. To usnadní jeho transformaci do podoby logického programu. K otestování funkčnosti programů budou použity rozsáhlé XML dokumenty, nad kterými budou formulovány testovací dotazy ověřující deduktivní schopnosti logického programu získaného transformací XML dokumentu. Předpokládáme provést vyhodnocování dotazů v prostředí logického jazyka Prolog a experimentálního deduktivního databázového systému. Důvodem jsou rozdílné omezující podmínky těchto dvou systémů. Získané výsledky z obou systémů budou porovnány.

Plánované indikátory dosažení - očekávané výsledky aktivity

Ověřená sada programů, kterou bude možné převést vstupní XML dokument do podoby logického programu.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Plánuje se otestování dané sady programů na rozsáhlých reálných datech. Vhodným kandidátem vstupních XML dokumentů může být např. databáze DBLP, kterou autor poskytuje ke stažení v XML formátu. Výsledky budou publikovány zejména na mezinárodních konferencích.

Číslo aktivity

13/10

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

SPOT – webová podpora on-line slovníku překladů odborných termínů

Zahájení aktivity

4.1.2010

Ukončení aktivity

30.6.2010

Popis aktivity

Cílem aktivity SPOT pro následující období bude plnohodnotná věcná a programová podpora pro překladatelské projekty a ověření možnosti sofistikované práce s taxonomií překladů jako základna pro výzkumné aktivity pracoviště.

Plánované indikátory dosažení - očekávané výsledky aktivity

Výsledkem aktivity bude

a) rutinní provoz slovníku s několika běžícími překladatelskými projekty,
b) veřejně přístupná část korpusu vzniklá na základě jejich činnosti,
c) uživateli průběžně upravovaná taxonomie termínů a jejich překladů používaná pro filtrování a cílené vyhledávání v korpusu slovníku.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Předpokládají se následující výstup aktivity:
Informace o počtu a aktivitě projektů dostupné z administrační části aplikace a převedené do formy statistik. Rozšířený korpus dostupný ve vyhledávací veřejné části aplikace spot.zcu.cz. Vylepšená taxonomie dostupná registrovaným uživatelům slovníku.
Příspěvek na konferenci a závěrečná výzkumná zpráva shrnující zkušenosti s uživatelskou prací nad společným překladovým korpusem.

Číslo aktivity

14/10

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

Ověřování, testování a diseminace systému WebGen

Zahájení aktivity

1.1.2010

Ukončení aktivity

31.12.2010

Popis aktivity

Testování a vylepšování postupů použitých při editaci stávajících prezentací. Návrh a vytvoření dialogového systému pro tvorbu nových designů webových prezentací. Testování dialogového vytváření nových designů pro webové prezentace.

Plánované indikátory dosažení - očekávané výsledky aktivity

Vylepšení dialogů pro editaci stavájících prezentací na základě zpětné vazby od uživatelů. Návrh a implementace dialogů pro vytváření nových designů webových prezentací s následným testováním uživateli.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Vylepšení dialogů pro editaci stavájících prezentací a dialogů pro vytváření nových designů webových prezentací.
Dosažené Výsledky budou prezentovány formou článků ve sbornících resp. technických zpráv v rámci FI Technical reports. Na stránkách laboratorře bude dostupná nová verze systému WebGen obsahující navržené postupy.

Číslo aktivity

15/10

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

Systém na generování grafiky na bázi grafických ontologií

Zahájení aktivity

1.1.2010

Ukončení aktivity

31.12.2010

Popis aktivity

Aktivita má za cíl vytvoření základních aplikací pro generování grafických objektů prostřednictvím dialogu. Prioritním cílem je umožnit nevidomým vytvářet elektronické grafické objekty a jednoduché grafické aplikace (např. Vánoční a Novoroční pozdrav apod.)

Plánované indikátory dosažení - očekávané výsledky aktivity

Bude vytvořena serverová aplikace založená na platformě
JavaEE a technologii Enterprise JavaBeans poskytující služby pro vytváření a vkládáná SvG grafiky, její anotaci za pomoci grafických ontologií, zpětné zíkávání popisu obrázků na základě jejich anotací a navigaci v SVG grafice pomocí rekursivní navigační mřížky. Budou vytvořeny grafické ontologie relevantní k základním aplikacím systému. Vytvořené aplikace budou předány k testování prostřednictví organizací podporujícíhc nevidomé a po úspěšném otestování budou v národním i mezinárodním měřítku diseminovány.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Vytvořený software, publikace ve sbornících mezinárodních konferencí.

Číslo aktivity

16/10

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

Kontroly a rozšiřování pokrytí slovesných valencí lexikonu VerbaLex na korpusových datech

Zahájení aktivity

1.1.2010

Ukončení aktivity

31.12.2010

Popis aktivity

Lexikon slovesných valencí VerbaLex již nyní reprezentuje nejrozsáhlejší lexikon tohoto typu pro češtinu. V minulém roce byly dokončeny několikanásobné kontroly konzistence lexikonu a jeho plné navázání na anglický Princetonský WordNet. V navazující aktivitě bude VerbaLex pomocí automatických metod syntaktické analýzy ověřován na velkém množství reálných korpusových a webových českých vět jednak ve smyslu určení (a zvýšení) pokrytí VerbaLexových slovesných rámců a jednak doplnění nových rámců podle reálných větných vzorů.

Plánované indikátory dosažení - očekávané výsledky aktivity

Nová verze lexikovu VerbaLex - plánujeme, že VerbaLex plně nahradí (a podstatně rozšíří) kompletní slovesnou část českého wordnetu

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

publikace

Číslo aktivity

17/10

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

Využití platformy DEB v aplikacích

Zahájení aktivity

1.1.2010

Ukončení aktivity

31.12.2010

Popis aktivity

Nástroje volně šířené platformy Dictionary Editor and Browser (DEB) mají široké využití, kde aktuálně vyvíjené systémy zahrnují aplikace pro Global WordNet Grid (vícejazyčná sémantická síť) a multiliguálně orientovanou lexikografickou stanici.

Plánované indikátory dosažení - očekávané výsledky aktivity

Nové verze nástrojů na platformě DEBII.
Doplňování a propojování sémantických sítí v dalších jazycích do nástroje DEBVisDic.

Nové vlastnosti nástroje DEBVisDic v rámci spolupráci v evropském projektu KYOTO.
Nový nástroj pro editaci slovníku anglických příjmení.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

publikace, nástroje

Číslo aktivity

18/10

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

Využití syntaktické a logické analýzy v aplikacích

Zahájení aktivity

1.1.2010

Ukončení aktivity

31.12.2010

Popis aktivity

Vlastní syntaktická a logická analýza textu poskytuje možnosti inteligentní analýzy textových dat pro široké spektrum aplikací. V rámci dané aktivity bude jednak dále vyvíjeno a vylepšováno jádro syntaktické a logické analýzy češtiny a jednak vzniknou nové pokročilé nástroje založené na aktuálních výstupech syntaktické analýzy.

Plánované indikátory dosažení - očekávané výsledky aktivity

Mezi plánované nové nástroje patří třeba nástroj na detekci a opravu interpunkce, nástroj na převod české věty na strukturu predikát-argumenty nebo nástroj na analýzu časových událostí ve větách umožňující využívat časové informace při vlastní analýze a vyvozování.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

publikace, nástroje

Číslo aktivity

19/10

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

Podobnost matematických textů ve vytvořeném korpusu

Zahájení aktivity

1.1.2010

Ukončení aktivity

31.12.2010

Popis aktivity

Vyhodnocování zpětné reakce algoritmů podobnosti matematických textů na http://dml.cz Usnadnění zpětné reakce na počítané podobnosti.

Plánované indikátory dosažení - očekávané výsledky aktivity

Rozšíření korpusu na 50000 článků a přepočítání podobností alespoň třemi metodami.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

uspořádání workshopu, nasazení podobnostního sw v reálném provozu dml.cz, publikace v odborném sborníku

---

Číslo aktivity

20/10

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

Extrakce lexikálních a funkčních vztahů z korpusových dat

Zahájení aktivity

1.1.2010

Ukončení aktivity

31.12.2010

Popis aktivity

Pro robustní zpracování přirozeného jazyka se ukazuje jako důležité použití lexikálních informací. Tyto informace o jednotlivých konkrétních slovech jsou dobře využitelné od morfologické analýzy přes syntaktickou analýzu až k sémantické analýze. Manuálně vytvářené zdroje bohužel nemají dostatečné pokrytí a přitom minimálně některé informace lze s dostatečnou spolehlivostí získávat automaticky z velmi velkých textových korpusů. Cílem aktivity je vytvoření algoritmů pro získávání takovýchto informací.

Plánované indikátory dosažení - očekávané výsledky aktivity

Algoritmy pro automatické získávání lexikálních informací o jednotlivých slovech či lemmatech a získávání funkčních vztahů mezi lemmaty z rozsáhlých korpusů. Algoritmy budou maximálním způsobem jazykově nezávislé a budou testovány na angličtině a češtině.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Publikace, automaticky získaná lexikální data

Číslo aktivity

21/10

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

Značkování anaforických vztahů v korpusových datech

Zahájení aktivity

1.1.2010

Ukončení aktivity

31.12.2010

Popis aktivity

Aplikace implementovaných metod pro automatickou analýzu anaforických vztahů na korpusová data.

Plánované indikátory dosažení - očekávané výsledky aktivity

Začlenění automaticky či poloautomaticky nalezených anaforických vztahů do vybraných korpusů obsažených v aktuální versi korpusového systému Manatee. Následné využití těchto dat vhodným doplněním fukncionality či visualisací v systému Bonito.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Publikace

3.2.2. NÁVRH ZMĚN V ŘEŠENÍ PROJEKTU - rok 2010

Pč.	Typ	Popis
1	návrh změn v řešení projektu	Žádná zásadní změna v řešení projektu není plánována, avšak v průběhu řešení se stejně jako v uplynulém roce mohou objevit další neplánované dílčí aktivity, které bude třeba vyřešit, aby mohlo být dosaženo plánovaných cílů projektu.

3.3. NÁKLADY PROJEKTU - rok 2010

3.3.1. NÁKLADOVÉ TABULKY ZA JEDNOTLIVÉ SUBJEKTY

Rok	2010
Typ	požadované
Organizace	Západočeská univerzita v Plzni
Role organizace	příjemce - koordinátor

POLOŽKA UZNANÝCH NÁKLADŮ tis. Kč	Náklady požadované tis. Kč	z toho požadované z účelové podpory tis. Kč
F1. - Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	**3115**	**2945**
F2. - Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	**0**	**0**
F3. - Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	**0**	**0**
F4. - Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	**80**	**0**
F5. - Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	**20**	**0**
F6. - Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	**170**	**0**
F7. - Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	**350**	**50**
F8. - Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, enegii a služby neuvedené výše	**330**	**0**
F9. CELKEM	**3995**	**2995**

	PŘEVOD DO fondu tis. Kč	POUŽITÍ Z fondu tis. Kč
F0. - Zúčtování s Fondem účelově určených prostředků	**0**	**70**

	ZDROJE FINANCOVÁNÍ CELKEM tis. Kč	- z toho Účelová podpora (DOTACE) tis. Kč	- z toho Ostatní veřejné zdroje tis. Kč	- z toho Neveřejné zdroje tis. Kč
Z9.	**3995**	**2995**	**0**	**1000**

Rok	2010
Typ	požadované
Organizace	Masarykova univerzita
Role organizace	příjemce

POLOŽKA UZNANÝCH NÁKLADŮ tis. Kč	Náklady požadované tis. Kč	z toho požadované z účelové podpory tis. Kč
F1. - Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	**1930**	**1777**
F2. - Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	**0**	**0**
F3. - Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	**34**	**0**
F4. - Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	**33**	**0**
F5. - Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	**0**	**0**
F6. - Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	**0**	**0**
F7. - Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	**180**	**0**
F8. - Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, enegii a služby neuvedené výše	**200**	**0**
F9. CELKEM	**2377**	**1777**

	PŘEVOD DO fondu tis. Kč	POUŽITÍ Z fondu tis. Kč
F0. - Zúčtování s Fondem účelově určených prostředků	**0**	**0**

	ZDROJE FINANCOVÁNÍ CELKEM tis. Kč	- z toho Účelová podpora (DOTACE) tis. Kč	- z toho Ostatní veřejné zdroje tis. Kč	- z toho Neveřejné zdroje tis. Kč
Z9.	**2377**	**1777**	**0**	**600**

3.3.2. NÁKLADOVÁ TABULKA ZA PROJEKT

Rok	2010
Typ	požadované
PROJEKT	2C06009 - CELKEM

POLOŽKA UZNANÝCH NÁKLADŮ tis. Kč	Náklady požadované tis. Kč	z toho požadované z účelové podpory tis. Kč
F1. - Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	**5045**	**4722**
F2. - Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	**0**	**0**
F3. - Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	**34**	**0**
F4. - Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	**113**	**0**
F5. - Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	**20**	**0**
F6. - Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	**170**	**0**
F7. - Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	**530**	**50**
F8. - Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, enegii a služby neuvedené výše	**530**	**0**
F9. CELKEM	**6372**	**4772**

	PŘEVOD DO fondu tis. Kč	POUŽITÍ Z fondu tis. Kč
F0. - Zúčtování s Fondem účelově určených prostředků	**0**	**70**

	ZDROJE FINANCOVÁNÍ CELKEM tis. Kč	- z toho Účelová podpora (DOTACE) tis. Kč	- z toho Ostatní veřejné zdroje tis. Kč	- z toho Neveřejné zdroje tis. Kč
Z9.	**6372**	**4772**	**0**	**1600**

3.3.3. NÁVRH ZMĚN V NÁKLADECH - rok 2010

Pč.	Typ	Popis
1	návrh změn v nákladech	Centrum ZPJ FI MU plánuje v roce 2010 dva přesuny ve skladbě finančních prostředků 1. navýšení mzdových prostředků o 53 tisíc. O stejnou částku budou sníženy provozní náklady projektu. Důvodem k této změně je navýšení mzdových tarifů MU v roce 2009 a zvýšení kvalifikace některých pracovníků. 2. z důvodu zjednodušení čerpání nákladů mezi jednotlivými interními zakázkami projektu (1 dotační a 3 nedotační zakázky) plánujeme celou částku z účelové podpory využít pouze na mzdové prostředky. Nejedná se v tomto případě o navýšení nebo změnu skladby celkových nákladů, jde pouze o přesun položkových nákladů mezi dotační a nedotační částí. Celkové dotační i nedotační náklady přitom zůstávají v souladu s původním plánem.

4. PŘÍLOHY

4.1. ZPRÁVA O POSTUPU ŘEŠENÍ PROJEKTU - rok 2009

4.1.1. POPIS ŘEŠENÍ PROJEKTU - seznam

	Pořadí	Soubor
	1	**Popis řešení projektu v roce 2009** Soubor obsahuje podrobný popis jednotlilvých aktivit naplněných při řešení projektu v průběhu roku 2009. Zprava2009_odst4_1_1.doc *(110 kB)*

4.1.2. DOSAŽENÉ VÝSLEDKY

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/01/2009**

Název výsledku

Ratio Statistics

Abstrakt

The results of independent experiments used for testing of certain method or certain phenomenon are often represented by samples, where ri, si are results from some statistically independent experiments. The main goal of our contribution is to analyse the random variable X = R/(R + S). The analysis is concerned with tolerance bounds of its statistic and introduces its asymptotic model as well. Furthermore, two particular cases of n are examined. In the first case, n is assumed to be non-random but changeable (e.g. comparison of automatic speech recognition methods). In the other case, n is assumed to be fixed (e.g. calculation of state unemployment rate from all district unemployment rates).

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AI, 2.- BD, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Umožňuje počítat toleranční intervaly úspěšnosti rozpoznávání slov, kde chyby ve slovech v jedné větě jsou statisticky závislé a tudíž není možné použít standardní modely založené na binomickém rozdělení.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Porovnání výsledné úspěšnosti modelů s ohledem na statistickou věrohodnost.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Pavelka Tomáš Ing. PhD.**
Spojení	377 632 491 377 632 402 tpavelka@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Vávra, F., Pavelka, T., Šedivá, B., Vokáčová, K., Marek, P., Neumanová, M.: Poměrové statistiky – toleranční intervaly, JČMF ROBUST 2008, Pribylina, Slovensko, 2008.	D – článek ve sborníku (RIV 2009)	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/02/2009**

Název výsledku

A Comparison of Acoustic Models Based on Neural Networks and Gaussian Mixtures

Abstrakt

This article tries to compare the performance of neural network and Gaussian mixture acoustic models (GMMs). We have carried out tests which match up various models in terms of speed and achieved recognition accuracy. Since the speed-accuracy trade-off is not only dependent on the acoustic model itself, but also on the settings of decoder parameters, we have suggested a comparison based on equal number of active states during the decoding search. Statistical significance measures are also discussed and a new method for confidence interval computation is introduced.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AI, 2.- BD, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Článek ukazuje praktické využití statistických metod popsaných ve výsledku "Ratio Statistics" v porovnání různých akustických modelů.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Výsledky ukazují, že pro určité typy problémů (klíčová je malá velikost slovníku) může použití neurovových sítí vést k lepší úspěšnosti rozpoznávání.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Pavelka Tomáš Ing. PhD.**
Spojení	377 632 491 377 632 402 tpavelka@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň liks.fav,zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Pavelka, T., Ekštein, K.: A Comparison of Acoustic Models Based on Neural Networks and Gaussian Mixtures. Proceedings of Text, Speech and Dialogue 2009, Plzeň, 2009	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/03/2009**

Název výsledku

Dialogue Act Recognition Approaches

Abstrakt

This paper deals with automatic dialogue act (DA) recognition. Dialogue acts are sentence-level units that represent states of a dialogue, such as questions, statements, hesitations, etc. The knowledge of dialogue act realizations in a discourse or dialogue is part of the speech understanding and dialogue analysis process. It is of great importance for many applications: dialogue systems, speech recognition, automatic machine translation, etc. The main goal of this paper is to study the existing works about DA recognition and to discuss their respective advantages and drawbacks. A major concern in the DA recognition domain is that, although a few DA annotation schemes seem now to emerge as standards, most of the time, these DA tag-sets have to be adapted to the specificities of a given application, which prevents the deployment of standardized DA databases and evaluation procedures. The focus of this review is put on the various kinds of information that can be used to recognize DAs, such as prosody, lexical, etc., and on the types of models proposed so far to capture this information. Combining these information sources tends to appear nowadays as a prerequisite to recognize DAs.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AI, 2.- BD, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Popis a srovnání existujících metod automatického rozpoznávání dialogových aktů. V práci jsou též popsána a diskutována různá anotační schémata.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Nové informace zjištěné při srovnání existujích anotačních schémat a metod automatického rozpoznávání dialogových aktů budou použity ke zvýšení přesnosti námi navržených a implementovaných metod, které modelují globální větnou strukturu (např. v kombinaci s lokální strukturou věty, apod.).

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Král Pavel Ing. PhD.**
Spojení	377 632 454 377 632 402 pkral@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Kral, P., Cerisara, C.: Dialogue Act Recognition Approaches, in Computing and Informatics, Slovenská akademie věd, 2010, (přijato k otištění 07/2009).		ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/04/2009**

Název výsledku

Evaluation of Automatic Speaker Recognition Approaches

Abstrakt

This paper deals with automatic speaker recognition in Czech. We focus here on context independent speaker recognition with a closed set of speakers. To the best of our knowledge, there is no comparative study about different speaker recognition approaches on the Czech language. The main goal of this paper is thus to evaluate and compare several parametrization/classification methods in order to build an efficient Czech speaker recognition system. All experiments are performed on a Czech speaker corpus that contains approximately half one hour of speech from ten Czech native speakers. Four parameterizations, which are mentioned in other studies as particularly successful for the speaker recognition task, are compared: Mel Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction Coefficients (PLPC), Linear Prediction Reflection Coefficients (LPREFC) and Linear Prediction Cepstral Coefficients (LPCEPSTRA). Two classifiers are compared: Hidden Markov Models (HMMs) and Multi-Layer Perceptron (MLP). In this work, we further study the impact of varying sizes of training corpus and test sentence on the recognition accuracy for different parametrizations and classifiers. For instance, we experimentally found that the recognition is still very accurate for test utterances as short as two seconds. The best recognition accuracy is obtained with LPCEPSTRA/LPREFC parametrizations and HMM classifier.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AI, 2.- BD, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Byla srovnána a vyhodnocena úspěšnost několika parametrizačních metod ve spojení se dvěma klasifikátory, skrytými Markovovovými modely (HMM) a vícevrstvým perceptronem (MLP), v úloze automatického rozpoznávání mluvčích.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Ukázali jsme, že zatímco MLP potřebuje méně trénovacích dat než HMM, skryté Markovovy modely dosahují o něco lepší přesnosti rozpoznávání než vícevrstvý percepron. Jako nejlepší parametrizační metody se v kombinaci s oběma klasifikátory ukázaly: kepstrální koeficienty lineární predikce (LPCEPSTRA) a reflexní koeficienty lineární predikce (LPREFC).

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Král Pavel Ing. PhD.**
Spojení	377 632 454 377 632 402 pkral@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Kral, P., Jezek, K., Jedlicka, P.: Evaluation of Automatic Speaker Recognition Approaches. In: 10th Western Pacific Acoustics Conference (WESPAC X 2009), Beijing, China, September 2009.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/05/2009**

Název výsledku

Nástroj pro srovnání výkonu relačních databází při zátěžových testech - jSQLBenchmark

Abstrakt

Aplikace jSQLBenchmark 1.0 slouží k jednoduchému testování relačních databázových systémů. Její metrikou je počet dotazů vykonaných za sekundu v závislosti na počtu paralelně přistupujících uživatelů. Aplikace disponuje grafickým uživatelským rozhraním, což činí práci s programem příjemnější. Veškeré parametry testu a parametry připojení k databázi se nastavují prostřednictvím k tomu určených komponent v uživatelském rozhraní, kde se rovněž zobrazují výsledy testu v grafické či tabulkové podobě.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- BD, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Hlavní výhodou nástroje jSQLBenchmark je jeho obecnost, tj. možnost připojení k libovolné databázi pomocí JDBC konektoru a nezávislost na testovacích datech. To je zajištěno tím, že aplikace nepoužívá žádnou interní (napevno zabudovanou) datovou strukturu, nad kterou se vykonávají předem specifikované testovací dotazy, ale struktudu databáze a sadu testovacích dotazů si určí uživatel.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Srovnali jsme výkon pěti nejrozšířenějších komerčních i volně dostupných databázových systémů (MySQL 5.0, PostgreSQL 8.1, Firebird 2.0, Oracle 10g a DB2 9). Ze srovnání vyplynulo, že v případě malých databází a nízkého počtu paralelně přistupujících uživatelů je vhodné použít databázový systém MySQL. V případě velké databáze (řádově několik milionů záznamů) a velkého počtu paralelně přistupujících uživatelů je nejvhodnější databázový systém Oracle.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Král Pavel Ing. PhD.**
Spojení	377 632 454 377 632 402 pkral@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	jSQLBenchmark, softwarový produkt, KIV FAV ZČU, Plzeň, 2009	R – software (RIV 2009)	
02	Čabrada, O., Král, P.: jSQLBenchmark Dokumentace, KIV FAV ZČU, Plzeň, 2009		CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/06/2009**

Název výsledku

Semantic Annotation for the LingvoSemantics Project

Abstrakt

In this paper, a methodology of semantic annotation of the LingvoSemantic corpus is presented. Semantic annotation is usually a time consuming and expensive process. We thus developed a methodology that significantly reduces the demands of the process. The methodology consists of a set of techniques and computer tools designed to simplify the process as much as possible. We claim that in this way it is possible to obtain sufficient amount of annotated data in a reasonable time frame. The LingvoSemantic project focuses on semantic analysis of user questions to an Internet information retrieval system. The semantic representation approach is based on abstract semantic annotation methodology. However, we advanced the annotation process. The bootstrapping method was used during the corpus annotation. The resulting annotated corpus consists of 20292 annotated sentences. In comparison to the straight-forward style of annotation, our approach significantly improved the efficiency of the annotation. The results, as well as a set of recommendations for creating the annotated data, are presented at the end of the paper.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AI, 2.- BD, 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

We have developed and evaluated a methodology for semantic annotation. It is based on a set of computer tools and techniques.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

The methodology proved to be efficient for annotating the corpora of 20k sentences. Moreover, the time demands have been rapidly decreased.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Habernal Ivan Ing.**
Spojení	377 632 456 377 632 402 habernal@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Habernal, I., Konopík, M.: Semantic Annotation for the LingvoSemantics Project. In: Text, speech and dialogue 2009. Berlin: Springer, 2009, s. 299-306. ISSN 0302-9743. ISBN 978-3-642-04207-2.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/07/2009**

Název výsledku

Hybrid Semantic Analysis

Abstrakt

This article is focused on the problem of meaning recognition in written utterances. The goal is to find a computer algorithm capable to construct the meaning description of a given utterance. An original system for meaning recognition is described in this paper. The key idea of the system is the hybrid combination of expert and machine-learning approaches to meaning recognition. The system utilizes a novel algorithm for semantic parsing. The algorithm is based upon extended context-free grammars. The grammars are automatically inferred from the data.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AI, 2.- BD, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

The testing of the system shows that the proposed architecture of a semantic analysis system can provide very good results. The architecture benefits from the advantages of both expert systems used for lexical class identification and from stochastic systems used for the parsing.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

The article describes an original system for semantic analysis.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Konopík Miloslav Ing. PhD.**
Spojení	377 632 456 377 632 402 konopik@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Konopík, M., Habernal, I.: Hybrid Semantic Analysis. In: Text, speech and dialogue 2009. Berlin: Springer, 2009, pp. 307-314. ISSN 0302-9743. ISBN 978-3-642-04207-2.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/08/2009**

Název výsledku

Context-dependent ASR

Abstrakt

Computer speech recognition gains more and more attention these days with its implementation in nearly everyday life. But the ultimate goal is still out of reach. The automatic recognition (ASR) systems can very precisely work on small domain. However the bigger the domain is the worse is the performance of the ASR system. The aim of many researchers is to diminish this problem on various levels of the ASR. This work describes components of an ASR system, how they are working together and delves into prosody and how it is used in ASR. From the usage of prosody, the main part of work describes how the ASR can be improved better modeling of the speech variance. We discuss usage of triphones, syllables and other models as well as algorithms and techniques for clustering.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AI, 2.- BD, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

We have successfully built and tested several triphone-based and syllable-based ASR systems. Thanks to context-dependency the baseline results of a syllable-based ASRs were much higher than a triphone ASR system when tested on spontaneous speech.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

This report describes an original and functional way how to improve ASR systems.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Hejtmánek Jan Ing.**
Spojení	377 632 458 377 632 402 hejtman2@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Hejtmánek, J.: PhD Study Report: Context-dependent ASR. Technical Report No. DCSE/TR-2009-12, August 2009, Pilsen		ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/09/2009**

Název výsledku

Stochastic Semantic Analysis

Abstrakt

This thesis presented overview of the current state-of-the-art methods for statistical semantic analysis. In comparison to the expert based systems, the main advantage of stochastic approaches is the ability to train the model from data. Furthermore, systems based on statistical models can be easily ported to other domains. However, the amount of the annotation effort must be also taken into account when developing the stochastic semantic analysis system. In this thesis, fundamental stochastic models along with the training and evaluation of these models are described.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AI, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

A comparison of various expert-based and stochastic semantic analysis approaches.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

An overview of the current state-of-the-art methods for statistical semantic analysis.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Habernal Ivan Ing.**
Spojení	377 632 456 377 632 402 habernal@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Habernal, I.: Stochastic Semantic Analysis. Technical report No. DCSE/TR-2009-04, Publisher: KIV ZČU, 2009		ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/10/2009**

Název výsledku

Web Services for Morphological Analysis

Abstrakt

The project "Web Services for Morphological Analysis" is aimed to provide a platform independent software library to process morphological analysis (morphological tagging) of the Czech language. It is written in Java and ANSI C language and it consists of server and client parts. The project is licensed under GPLv3 licence. Detailed and up-to-date documentation can be found at http://liks.fav.zcu.cz/mediawiki/index.php /WebServicesForMorphologicalAnalysis

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- BD, 2.- AI, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Platform independent software library to process morphological analysis (morphological tagging) of the Czech language.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Software library allows to incorporate morphological analysis into various clients.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Habernal Ivan Ing.**
Spojení	377 632 456 377 632 402 habernal@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	http://liks.fav.zcu.cz/mediawiki/index.php /WebServicesForMorphologicalAnalysis	GA – Prototyp (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/11/2009**

Název výsledku

LINGVO Annotation Manager

Abstrakt

Aplikace LINGVO Annotation Manager (LAM) slouží ke správě a k ukládání velkého množství sémantických dat. LAM ukládá data v široce používaném formalismu nazývaným abstraktní sémantické anotace. Známý formalismus umožňuje využít aplikaci dalšími subjekty, které se sémantickými anotacemi zabývají. LAM je důkladně otestován, jelikož se v našich projektech používá více než tři roky. Během této doby byl manažer použit k uložení více než 20 000 vět.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AI, 2.- BD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Celá aplikace je založena na nové metodologii anotování dat. Použití aplikace umožní systematický přístup k procesu anotování dat.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Aplikace umožní snížit náklady potřebné pro sémantické anotování dat a umožní celý proces zrychlit a zpřehlednit.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Konopík Miloslav Ing. PhD.**
Spojení	377 632 456 377 632 402 konopik@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	https://liks.fav.zcu.cz/mediawiki/index.php/LAM	R – software (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/12/2009**

Název výsledku

Hybrid Semantic Analysis

Abstrakt

Disertační práce se věnuje vývoji originálního systému pro sémantickou analýzu. Systém je vyvíjen pro český jazyk v kontextu aplikace inteligentního vyhledávání na internetu. Značná část práce se zabývá procesem získávání trénovacích dat (tzv. sémantické anotování). Je zde popsána rychlá a spolehlivá metoda, jak tato data získat. Základní myšlenky vytvořeného systému jsou založeny na systému Chronos a na HVS modelu. Vyvinutý přístup k sémantické analýze používá hybridní kombinaci expertních metod a metod strojového učení. Při vývoji systému byl vytvořen nový algoritmus pro sémantické parsování. Algoritmus používá aktivní chart parser a rozšířené bezkontextové gramatiky. Práce ukazuje, že hybridní kombinace algoritmů představuje způsob jak vytvořit robustní systém pro sémantickou analýzu.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AI, 2.- JC, 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Práce obsahuje dva základní inovační aspekty. Prvním je nový algoritmus pro parsování vět přirozeného jazyka. Druhým je vylepšená metodologie pro snazší sémantické anotování vět.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Navržené metody pro automatické zpracování sémantické informace umožňují přesnější zpracování sémantiky v počítači.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Konopík Miloslav Ing. PhD.**
Spojení	377 632 456 377 632 402 konopik@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Konopík, M.: Hybrid Semantic Analysis, Doctoral Thesis, Pilsen, 2009.		ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/13/2009**

Název výsledku

LTranscriber - autorizovaný software

Abstrakt

LTranscriber je podpůrná aplikace pro systém LASER/LINGVO, která umožňuje a usnadňuje pořizování přepisu nahrávek ze zvukového korpusu. Pořízení korpusu je zásadní podmínkou úspěšného natrénování rozpoznávače přirozené řeči. Korpus se skládá z nahrávek a jejich přepisů. Software LTranscriber usnadňuje velmi namáhavý ruční přepis zvukového záznamu do ortografické nebo kombinované ortograficko-fonetické podoby. Aplikace umožňuje rozčlenit záznam na kratší úseky, dovoluje vymezování rozsahu jednotlivých pasáží přepisu, dovoluje pracovníkovi transkripce zpomalovat přehrávání záznamu, pozastavovat, opakovat neustále vybraný úsek, velmi jednoduchým způsobem upravovat hranice segmentů záznamu, zoomovat, atd. Práci pracovníka transkripce urychluje také tím, že např. vkládání speciálních fonetických značek se děje výběrem z inteligentní nabídky, která se automaticky adaptuje podle právě prováděné úlohy. Aplikace se dále stará o ukládání přepisu v přenositelném formátu založeném na XML.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- JC, 2.- AI, 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Oproti existujícímu software (např. Transcriber 1.5.1) je LTranscriber výrazně modernější - naprogramovaný v jazyce Java 1.6 a tedy naprosto přenositelný mezi různými operačními prostředími. Také uživatelské rozhraní je pro pracovníka pořizujícího transkripci mnohem příjemnější a zvyšuje jeho efektivitu přidáním inteligentních asistenčních funkcí, zejména možnost zoomu audiosignálu, opakování segmentu s možností posunu hranic apod.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Podstatné zvýšení výkonu specialistů pořizujících transkripty korpusových záznamů oproti stavu při používání starších aplikací.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Ekštein Kamil Ing. PhD.**
Spojení	377 632 406 377 632 402 kekstein@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Dokumentace k software je jeho součástí.		ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/14/2009**

Název výsledku

Konferenční portál TSD - autorizovaný software

Abstrakt

Vyvinutý software Konferenční portál TSD byl původně zamýšlen jako komplexní řešení agendy konference Text, Speech and Dialogue 2009, ovšem díky vysoce modulární architektuře a důslednému oddělení prezentační vrstvy od aplikační logiky je možné jej velice snadno nasadit na libovolné jiné akce, které rozsahem a složitostí agendy odpovídají mezinárodní konferenci TSD. Portál podporuje kromě správy webových stránek konference ve stylu CMS (Content Management System) také komunikaci s registrovanými účastníky jednak pomocí osobních kont a jednak také konfigurovatelným hromadným mailerem. Dále portál zajišťuje registraci účastníků konference, sběr jejich příspěvků, anonymní nezávislou recenzi příspěvků několika (obecně N) recenzenty, shromáždění camera-ready verzí přijatých příspěvků, přípravu dat k sestavení sborníků, evidenci došlých plateb konferenčního poplatku, aj. Software pracuje na platformě LAMP (Linux, Apache, MySQL, PHP), je velmi snadno přenositelný, rekonfigurovatelný a doplnitelný o další potřebné moduly.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- JC, 2.- JD, 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Vyvinutý software vhodným způsobem kombinuje vlastnosti generického CMS s řešením webové služby na míru, čímž zajišťuje z uživatelského i administrátorského hlediska velmi jednoduché, snadno ovladatelné, rychlé a příjemné prostředí, které obvykle u řešení toho typu chybí (portály jsou vesměs značně nepřehledné a komplikované).

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Značné usnadnění přípravy konference, omezení zátěže organizátorů administrativními úkony, významné omezení možnosti zanesení chyby do agendy konference, předcházení problémům typu opomenutí zařazení článku do sborníku, apod. Přínosné je také zprůhlednění procesu výběru článků (portál neumožňuje "obcházet" recenzní řízení) a spolehlivá evidence příjmů konference z účastnických poplatků.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Ekštein Kamil Ing. PhD.**
Spojení	377 632 406 377 632 402 kekstein@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň http://liks.fav.zcu.cz, http://www.kiv.zcu.cz/tsd2009

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Dokumentace přiložená k softwaru nemá konkrétní název, jedná se o webovou aplikaci umožňující snadné použití zpracovaného softwaru.		ANG
02	Homolka, J.: Software pro prezentaci a správu vědeckých konferencí. Bakalářská práce, KIV FAV ZČU, Plzeň, 2009.		CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/15/2009**

Název výsledku

WEBSOM method - word categories in Czech written documents

Abstrakt

We applied well-known WEBSOM method (based on two layer architecture) to categorization of Czech written documents. Our research was focused on the syntactic and semantic relationship within word categories of word category map (WCM). The document classification system was tested on a subset of 100 documents (manual work was necessary) from the corpus of Czech News Agency documents. The result confirmed that WEBSOM method could be hardly evaluated because humans have problems with natural language semantics and determination of semantic domains from word categories.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- JC, 2.- BD, 3.- AI, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Aplikace metody WEBSOM na česky psané dokumenty a rozsáhlé testy detekce sémantických domén (provedeny na dokumentech ČTK).

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Testy potvrdily, že je obtížně realizovatelné stanovit metodiku ohodnocení výsledků aplikace metody WEBSOM, protože testované subjekty měly problém s určením sémantických domén.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Mouček Roman Ing. PhD.**
Spojení	377 632 465 377 632 402 moucek@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Mouček R., Mautner P.: WEBSOM method - word categories in Czech written documents. In: Text, speech and dialogue 2009. Berlin: Springer, 2009. s. 85-92. ISSN 0302-9743. ISBN 978-3-642-04207-2.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/16/2009**

Název výsledku

Sémantika přirozeného jazyka

Abstrakt

Poisované téma se zabývá možnostmi počítačového zpracování sémantiky přirozeného jazyka a reálného světa a pokládá otázku, do jaké míry je toto zpracování možné a smysluplné. Odpověď pak hledá v kombinaci poznatků a zkušeností tří různých oborů – neurověd, lingvistiky a informatiky. Stručně je prezentován pohled neurověd na fungování lidského mozku a popsány paměťové složky mající vliv na zpracování sémantiky přirozeného jazyka a sémantiky vnějšího reálného světa. Krátce je představen i vnější, lingvistický pohled na přirozený jazyk a jeho sémantické roviny. Z informatických oborů jsou pak představeny přístupy umělé inteligence a softwarového inženýrství. Zmíněna je i vize sémantického webu.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AI, 2.- BD, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Počítačové zpracování sémantiky přirozeného jazyka zahrnuje schopnost počítačového systému smysluplně interpretovat text či promluvu v přirozeném jazyce a posléze na tento text či promluvu adekvátně reagovat podrobně posána je schopnost počítačového „porozumění“ přirozenému jazyku.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Vývoj nového modelovacího prostředku pro popis a porozumění přirozenému jazyku - prostředku s výpočetní kapacitou, uspořádáním a fungováním obdobným funkci lidského mozku. Dále jde o popis vize sémantického webu, která může určovat hranici, k níž z pohledu zpracování sémantiky jazyka a reálného světa stačí v informatice dojít.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Matoušek Václav Prof. Ing. CSc.**
Spojení	377 632 471 377 632 402 matousek@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni ulniverzitní 2732 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Matoušek, V., Mouček, R., Mautner, P.: Sémantika přirozeného jazyka a reálného světa - počítačové zpracování. In: Hlaváčková, D., Horák, A., Osolsobě, K., Rychlý P. (Eds.): After Half a Century of Slavonic Natural Language Processing. Masaryk University, Brno, 2009. ISBN 978-80-7399-815-8.	C – kapitola v odborné knize (RIV 2009)	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/17/2009**

Název výsledku

SUTLER: Update SummarizER based on Latent Topics

Abstrakt

This paper deals with our past and recent research in text summarization. We went from single-document summarization through multi-document summarization to update summarization. We describe the development of our summarizer which is based on latent semantic analysis (LSA). The classical LSA-based summarization model was improved by Iterative Residual Rescaling. We propose the update summarization component which determines the redundancy and novelty of each topic discovered by LSA. Moreover, we have modified the sentence selection component in order to prevent inner summary redundancy. The results of our first participation in TAC/DUC evaluation seem to be promising.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- BD, 2.- JD, 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Aktualizační sumarizace novinových článků založená na LSA.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Nová metoda aktualizační sumarizace.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Steinberger Josef Ing. PhD.**
Spojení	377 632 461 377 632 402 jstein@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň www.kiv.zcu.cz/vyzkum

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Steinberger J., Ježek K.: SUTLER: Update Summarizer Based on Latent Topics. In Proceedings of TAC"08, NIST, Gaithersburgh, United States, 2009. http://www.nist.gov/tac/publications/		ANG
02	Steinberger J.: Aktualizační sumarizace textů (in Czech). Proceedings of Znalosti 2009, Brno, Czech Republic, February 2009, pp. 234–245, ISBN 978-80-227-3015-0.	D – článek ve sborníku (RIV 2009)	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/18/2009**

Název výsledku

Text Summarization: An Old Challenge and New Approaches

Abstrakt

One of the most relevant today's problems called "information overloading" has increased the necessity of more sophisticated and powerful information compression methods - summarizers. This chapter firstly introduces taxonomy of summarization methods, an overview of their principles from classical ones, over corpus based, to knowledge rich approaches. We consider various aspects which can affect their categorization. A special attention is devoted to application of recent information reduction methods, based on algebraic transformations. Our own LSA (Latent Semantic Analysis) based approach is included too. The next part is devoted to evaluation measures for assessing quality of a summary. The taxonomy of evaluation measures is presented and their features are discussed. Further, we introduce experiences with the development of our web searching and summarization system. Finally, some new ideas and a conception for the future of this field are mentioned.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- BD, 2.- JD, 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Původní srovnávací studie moderních metod sumarizace textů, včetně zkušeností z vývoje a testování naší vlastní nové sumarizační metody.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Souhrnná studie existujících metod sumarizace textů, včetně zkušeností z vývoje a testování naší vlastní nové sumarizační metody.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Steinberger Josef Ing. PhD.**
Spojení	377 632 461 377 632 402 jstein@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň www.kiv.zcu.cz/vyzkum/

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Steinberger J., Jezek K.: Text Summarization: An Old Challenge and New Approaches. In: Foundations of Computational Intelligence Vol.6, pages 127- 149, Data Mining Book Series, Springer, ISSN 1860-949X, ISBN 978-3-642-01090-3, 2009	C – kapitola v odborné knize (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/19/2009**

Název výsledku

Evaluation Measures for Text Summarization

Abstrakt

We explain the ideas of automatic text summarization approaches and the taxonomy of summary evaluation methods. Moreover, we propose a new evaluation measure for assessing the quality of a summary. The core of the measure is covered by Latent Semantic Analysis (LSA) which can capture the main topics of a document. The summarization systems are ranked according to the similarity of the main topics of their summaries and their reference documents. Results show a high correlation between human rankings and the LSA-based evaluation measure. The measure is designed to compare a summary with its full text. It can compare a summary with a human written abstract as well however, in this case using a standard ROUGE measure gives more precise results. Nevertheless, if abstracts are not available for a given corpus, using the LSA-based measure is an appropriate choice.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- BD, 2.- JD, 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Přehled metod hodnotících kvalitu souhrnů, popis metody založené na LSA.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Nová metoda hodnocení kvality souhrnů

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Steinberger Josef Ing. PhD.**
Spojení	377 632 461 377 632 402 jstein@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň www.kiv.zcu.cz/výzkum/

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Steinberger J., Ježek K.: Evaluation Measures for Text Summarization. Computing and Informatics, volume 28 (2009), number 2, pages 251-275, Slovak Academy of Sciences, ISSN 1335-9150.	J – článek v odborném periodiku (časopise) (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/20/2009**

Název výsledku

Update Summarization Based on Latent Semantic Analysis

Abstrakt

This paper deals with our recent research in text summarization. We went from single-document summarization through multi-document summarization to update summarization. We describe the development of our summarizer which is based on latent semantic analysis (LSA) and propose the update summarization component which determines the redundancy and novelty of each topic discovered by LSA. The final part of this paper presents the results of our participation in the experiment of Text Analysis Conference 2008.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- BD, 2.- JD, 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Popis metody aktualizační sumarizace založené na LSA.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Nová metoda aktualizační sumarizace

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Steinberger Josef Ing. PhD.**
Spojení	377 632 461 377 632 402 jstein@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň www.kiv.zcu.cz/vyzkum/

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Steinberger J., Jezek K.: Update Summarization Based on Latent Semantic Analysis. Proceedings of 12th International Conference, TSD 2009, Pilsen, Czech Republic, September 2009. LNAI 5729, Springer-Verlag Berlin Heidelberg New York, ISBN 978-3-642-04207-2, ISSN 0302-9743.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/21/2009**

Název výsledku

Update Summarization Based on Novel Topic Distribution

Abstrakt

This paper deals with our recent research in text summarization. The field has moved from multi-document summarization to update summarization. When producing an update summary of a set of topic-related documents the summarizer assumes prior knowledge of the reader determined by a set of older documents of the same topic. The update summarizer thus must solve a novelty vs. redundancy problem. We describe the development of our summarizer which is based on Iterative Residual Rescaling (IRR) that creates the latent semantic space of a set of documents under consideration. IRR generalizes Singular Value Decomposition (SVD) and enables to control the influence of major and minor topics in the latent space. Our sentence-extractive summarization method computes the redundancy, novelty and significance of each topic. These values are finally used in the sentence selection process. The sentence selection component prevents inner summary redundancy. The results of our participation in TAC evaluation seem to be promising.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- BD, 2.- JD, 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Popis metody aktualizační sumarizace založené na LSA/IRR.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Nová metoda aktualizační sumarizace.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Steinberger Josef Ing. PhD.**
Spojení	377 632 461 377 632 402
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň www.kiv.zcu.cz/vyzkum/

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Steinberger J., Jezek K.: Update Summarization Based on Novel Topic Distribution. In Proceedings of the 2009 ACM Symposium on Document Engineering, Munich, Germany, September 2009. Association for Computing Machinery, ISBN 978-1-60558-575-8.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/22/2009**

Název výsledku

Multilingual Statistical News Summarisation: Preliminary Experiments with English

Abstrakt

In this paper we present a generic approach for summarising multilingual news clusters such as the ones produced by the Europe Media Monitor (EMM) system. It is generic because it uses robust statistical techniques to perform the summarisation step and its multilinguality is inherited from the multilingual entity disambiguation system used to build the source representation. We ran preliminary experiments with the TAC 2008 data, an English corpus for summarization research, and we obtained promising improvements over a summarisation system ranked in the top 20% at the TAC 2008 competition

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- BD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Využití rezoluce jmen pro sumarizaci textů.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Vylepšení sumarizační metody

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Steinberger Josef Ing. PhD.**
Spojení	377 632 461 377 632 402 jstein@kiv.zcu.cz
Organizace	49777513 Západočeská univrzita v Plzni Univerzitní 2732 8 30614 Plzeň www.kiv.zcu.cz/vyzkum/

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Kabadjov M. A., Steinberger J., Pouliguen B., Steinberger R., Poesio M.: Multilingual Statistical News Summarisation: Preliminary Experiments with English. Proceedings of the workshop ""Intelligent Analysis and Processing of Web News Content"" (WI-IAT""09). Milano, Italy, IEEE-CS Press, September 2009. ISBN 978-0-7695-3801-3.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/23/2009**

Název výsledku

ALMUS: Automatický sumarizátor textů

Abstrakt

Automatický sumarizátor textů je založen na latentní sémantické analýze. Vytváří základní souhrn ze starších dokumentů a aktualizační souhrn z nových dokumentů pro dané téma. Systém byl použit při tvorbě souhrnů pro TAC 08. Více informací v http://www.nist.gov/tac/publications/2008/participant.papers/Sutler.proceedings.pdf.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Sumarizace textů metodou založenou na LSA.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Použit při tvorbě souhrnů pro TAC08.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Steinberger Josef Ing. PhD.**
Spojení	377 632 461 377 632 402 jstein@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň www.kiv.zcu.cz/vyzkum/software

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Dokumentace je dostupná na http://www.kiv.zcu.cz /vyzkum/software/	R – software (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/24/2009**

Název výsledku

An Experience with Building Digital Open Access Repository DML-CZ

Abstrakt

The growth of digital repositories of scientific documents is speed-ed up by various digitisation activities. Almost all papers of mathematical journals are reviewed by either Mathematical Reviews or ZentralBlatt Math, summing up to more than 2.000.000 entries. In the paper we discuss possibilities and experiments we did on the data of Czech Digital Mathematics Library, DML-CZ with the goal of developing novel scalable methods of document classification and retrieval of multilingual mathematical papers.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- AF, 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Metadata ukládaná společně s dokumenty umožňují lepší a přesnější vyhledávání v digitální matematické knihovně.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

The Czech Digital Mathematics Library (DML-CZ) has been developed in order to preserve in a digital form the content of major part of mathematical literature that has ever been published in the Czech lands, and to provide a free access to the digital content and bibliographical data.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Sojka Petr doc. RNDr. Ph.D.**
Spojení	+420 549 494 377 sojka@mail.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Sojka, Petr. An Experience with Building Digital Open Access Repository DML-CZ. In Proceedings of CASLIN 2009, Institutional Online Repositories and Open Access, 16th International Seminar. první. Pilsen, Czech Republic : University of West Bohemia, Pilsen, 2009. od s. 74-78, 5 s. ISBN 978-80-7043-806-0.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/25/2009**

Název výsledku

Building of Corpus Based E-learning Materials for Czech

Abstrakt

The quality of data plays a crucial role in the effectiveness of the (e-)learning process. Many e-learning materials lack connection with real-world examples and the preparation of quality materials is a highly time-consuming work for domain professionals. We present a corpus based method for preparing e-learning tools which are used for education of Czech in a particular subject. The used corpus has been created in past years from the texts written by previous students of this subject. Moreover, the corpus is annotated with mistakes and their corrections, categorized according to a linguistic and typographic classification, and hence it is possible to obtain mistake-specific real-world examples. We also present a specific usage in the information system of Masaryk University in exercises related to punctuation in Czech sentences.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Využití ručně značkovaného korpusu chyb k automatickému budování e-learningových materiálů pro výuku českého jazyka.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Využití ručně značkovaného korpusu chyb umožňuje rychlejší, přesnější vytváření e-learningových materiálů, které více odpovídají praktickému užití jazyka.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Bušta Jan Bc.**
Spojení	+420 549 494 377 malpa@mail.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Bušta, Jan - Jakubíček, Miloš. Building of Corpus Based E-learning Materials for Czech. In SCO 2009 : Sharable Content Objects : 6. ročník konference o elektronické podpoře výuky. 1. vyd. Brno : Masarykova univerzita, 2009. od s. 144-149, 6 s. ISBN 978-80-210-4878-2.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/26/2009**

Název výsledku

Classification of Errors in Text

Abstrakt

This paper presents two classifications of errors in Czech texts. As a basic resource we use the corpus (Chyby Errors) which has been continuously developed from 19992000. The corpus text contains various kinds of errors such as spelling, typographical, grammatical, semantic, lexical, and stylistic ones. They have been corrected manually and annotated according to the classification of errors (annotation scheme) developed for this purpose. For the annotation we implemented a tool named WinCorr. We mention the first annotation scheme and discuss the second one which has been designed recently to obtain more adequate description of the errors occurring in texts. We also discuss the principles on which both classifications are based.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Klasifikace pomocí nástroje WinCorr je maximálně jednoznačná a rychlá. Usnadňuje ruční rozlišení různých typů chyb v korpusu chyb.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Z ručně anotovaných a klasifikovaných dat z textového korpusu chyb je možné automaticky vytvářet například testy, praktické příklady chyb atd., které jsou založeny na skutečném jazyce.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Hlaváčková Dana Mgr. Ph.D.**
Spojení	+420 549 491 864 17907@mail.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Jakubíček, Miloš - Bušta, Jan - Hlaváčková, Dana - Pala, Karel. Classification of Errors in Text. In RASLAN 2009 : Recent Advances in Slavonic Natural Language Processing. 1. vyd. Brno : Masaryk University, 2009. od s. 109-119, 11 s. ISBN 978-80-210-5048-8.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/27/2009**

Název výsledku

Cornetto Tools and Methodology for Interlinking Lexical Units, Synsets and Ontology

Abstrakt

Cornetto project produces a lexical semantic database for Dutch. The database combines Wordnet with FrameNet-like information. The data are derived from two existing lexical resources: the Dutch Wordnet and the Referentie Bestand Nederlands. For storing and editing this complex database, we used the Dictionary Editor and Browser platform.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Cornetto svazuje data z WordNetu a FrameNetu (s využitím platformy DEB) a vytváří tak nový, bohatší zdroj lexikálních dat.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Díky provázání dat z WordNetu a FrameNetu umožňuje Cornetto využít zároveň výhod obou zdrojů (synsety, ontologie a lexikální jednotky).

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Horák Aleš RNDr. Ph.D.**
Spojení	+420 549 494 377 hales@mail.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Horák, Aleš - Rambousek, Adam - Vossen, Piek - Segers, Roxane - Maks, Isa - van der Vliet, Hennie. Cornetto Tools and Methodology for Interlinking Lexical Units, Synsets and Ontology. In Current Issues in Unity and Diversity of Languages. Seoul, Republic of Korea : The Linguistic Society of Korea, 2009. od s. 2695-2713, 19 s. ISBN 978-89-90696-71-7.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/28/2009**

Název výsledku

Czech Word Sketch Relations with Full Syntax Parser

Abstrakt

This paper describes the exploitation of dependency relations obtained from syntactic parsing of Czech for building new Czech Word Sketch tables. Standard Word Sketch construction process usually uses so called Sketch grammars a simplified process of identifying dependency relations based on regular expressions. This may, of course, lead to errors, which should however not influence (so much) the overall numbers computed on a very big corpus. The paper presents an experiment of using relations resulting from full syntactic parsing will they perform better than the standard Sketch grammar or not?

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Využití syntaktického analyzátoru pro zpracování Word Sketchů (namísto Sketch gramatik založených na regulárních výrazech).

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Využití syntaktického analyzátoru může obohatit a případně zvýšit kvalitu výsledných Word Sketchů.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Horák Aleš RNDr. Ph.D.**
Spojení	+420 549 494 377 hales@mail.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Horák, Aleš - Rychlý, Pavel - Kilgarriff, Adam. Czech Word Sketch Relations with Full Syntax Parser. In After Half a Century of Slavonic Natural Language Processing. Brno, Czech Republic : Masaryk University, 2009. od s. 101-112, 12 s. ISBN 978-80-7399-815-8.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/29/2009**

Název výsledku

Digitisation Workflow in the Czech Digital Mathematics Library

Abstrakt

Experience in setting up a workflow from scanned images of mathematical writings into a fully fledged mathematical library is described on the example of the project Czech Digital Mathematics Library DML-CZ http://dml.cz. An overview of the whole process is given, with detailed description of production steps involving scanned image processing and optical character recognition. Experience gained, lessons learned and tools prepared during development of DML-CZ are described. DML-CZ now serves over 25,600 articles (275,000 digitised pages) to the public.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Proces digitalizace matematických textů zahrnuje mimo jiné přidání metadat ke každému článku, na jejichž základě lze obsahově porovnávat jednotlivé články a zvyšovat tak pohodlí při fulltextovém vyhledávání v knihovně.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Česká digitální matematická knihovna DML-CZ obsahuje přes 25 tisíc kompletně digitalizovaných článků z oblasti matematiky, dovoluje fulltextové vyhledáván, vyhledávání podobných článků atd.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Sojka Petr doc. RNDr. Ph.D.**
Spojení	+420 549 494 377 sojka@mail.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Sojka, Petr. Digitisation Workflow in the Czech Digital Mathematics Library. Math-for-Industry, Kyushu, Japan : Faculty of Mathematics,Kyushu University, 2009, 22, od s. 272-280, 9 s. ISSN 1881-4042. 2009.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/30/2009**

Název výsledku

Discovering Grammatical Relations in Czech Sentences

Abstrakt

The syntactic parser synt developed at NLP Centre, Faculty of Informatics, Masaryk University, can provide as one of its possible outputs a list of dependency relations discovered in the analysed sentence. In the paper, we present the result of codification and translation of the (rather technically labeled) dependency relations from synt to linguistically significant relations. The resulting relations are demonstrated by means of Word Sketches (WS), where the new relations are compared with traditional WS relations from WS grammar.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Jeden z možných výstupů syntaktického analyzátoru synt - seznamy závislostních vztahů - je použit pro vytvoření Word Sketchů, které jsou obvykle budovány na základě Sketch gramatik.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Využití daného výstupu ze syntaktického analyzátoru synt dovoluje obohatit a zvýšit kvalitu výsledných Word Sketchů.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Horák Aleš RNDr. Ph.D.**
Spojení	+420 549 494 377 hales@mail.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Horák, Aleš - Rychlý, Pavel. Discovering Grammatical Relations in Czech Sentences. In Proceedings of the RASLAN Workshop 2009. první. Brno : Masaryk University, 2009. od s. 81-90, 9 s. ISBN 978-80-210-5048-8.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/31/2009**

Název výsledku

Exploring and Extending Czech WordNet and VerbaLex

Abstrakt

This paper presents usage of two major, linguist-made lexical resources of Czech language: WordNet and VerbaLex. First, a conversion to RDF was made. Afterwards, a Prolog program was used to analyse Czech language inputs. In the second part of the article an extension to current VerbaLex is proposed. Possible pitfalls are discussed. In the conclusion, we emphasize the side-effect of this work: an important feedback for authors and administrators of both lexical resources.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Analýza českých vět pomocí kombinace dvou významných datových zdrojů: WordNetu a VerbaLexu.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

WordNet poskytuje lexikální data a VerbaLex valenční slovesné rámce, které v kombinaci umožňují efektivně analyzovat české věty.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Nevěřilová Zuzana Mgr.**
Spojení	+420 549 494 377 xpopelk@mail.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Nevěřilová, Zuzana. Exploring and Extending Czech WordNet and VerbaLex. In Proceedings of the RASLAN Workshop 2009. 1. vyd. Brno : Masaryk University, 2009. 6 s. ISBN 978-80-210-5048-8.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/32/2009**

Název výsledku

Exploring Derivational Relations in Czech with the Deriv Tool?

Abstrakt

The aim of this paper is to present a tool for testing automatic word derivation for Czech. The derivation of some word formation types in Czech is in high degree regular. It can be described by formal rules. A new version of the web interface Deriv working with the morphological analyzer ajka enables us to formulate more complex word formation rules and to test more complicated cases of derivational relations. The second issue touched extensively in the paper are the types of derivational relations and their semantic classification. We have proposed 14 semantic classes for suffixes and 11 for prefixes. The tool Deriv helps considerably in establishing semantics of the derivational relations.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Webové rozhraní nástroje Deriv spolupracující s morfologickým analyzátorem češtiny ajka výrazně usnadňuje návrh a správu sémantiky derivačních vztahů.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Nástroj Deriv poskytuje prostředí pro testování mezí a možností automatického vyhledávání derivačních vztahů v češtině.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Pala Karel doc. PhDr. CSc.**
Spojení	+420 549 495 616 pala@fi.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Osolsobě, Klára - Hlaváčková, Dana - Pala, Karel - Šmerk, Pavel. Exploring Derivational Relations in Czech with the Deriv Tool? In NLP, Corpus Linguistics, Corpus Based Language Research. Bratislava, Slovakia : Tribun, 2009. od s. 152-161, 10 s. ISBN 978-80-7399-875-2.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/33/2009**

Název výsledku

Fast Morphological Analysis of Czech

Abstrakt

Paper presents a new Czech morphological analyser which takes an advantage of Jan Daciuk""s algorithms for minimal deterministic acyclic finite state automata. The new analyser is six times faster than the current analyser ajka concerning the proper analysis, i.e. returning possible lemmata and tags for a given word form, but for some other related tasks is the difference even bigger.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Využití výhod algoritmu minimálního deterministického acyklického konečného automatu pro morfologickou analýzu češtiny.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Nový algoritmus pracuje minimálně 6x rychleji než dosavadní algoritmus morfologické analýzy.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Šmerk Pavel RNDr.**
Spojení	+420 549 494 347 smerk@fi.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Šmerk, Pavel. Fast Morphological Analysis of Czech. In Proceedings of the Raslan Workshop 2009. Brno : Masarykova univerzita, 2009. 4 s. ISBN 978-80-210-5048-8.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/34/2009**

Název výsledku

Integrální počet funkcí více proměnných

Abstrakt

Interaktivní elektronická sbírka příkladů Integrální počet funkcí více proměnných je určena pro podporu výuky Integrálního počtu funkcí více proměnných v předmětech Matematická analýza IV (M5520), Matematická analýza III (M3100), Matematika III (MB103). Obsahuje interaktivní 3D grafiku a testové otázky k ověření pochopení probírané problematiky. Dokument je vysázen pdfLaTeXem, pro tvorbu testů byl využit volně šiřitelný latexovský balíček AcroTeX, 3D obrázky byly vkládány ve formátu U3D. Sbírka je realizována v PDF formátu a je tedy nezávislá na platformě. Pro zobrazení plně funkční publikace je nezbytné mít na počítači nainstalovaný Adobe Reader 8.1.1 nebo novější. Sbírka byla na Konferenci a soutěži Elearning Hradec Králové (10.-12. 11. 2009) oceněna Cenou České asociace distančního univerzitního vzdělávání.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- BA, 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Interaktivní elektronická sbírka příkladů obsahuje interaktivní 3D grafiku a je k dispozici ve formátu PDF nezávislém na platformě uživatelů.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Sbírka obsahuje interaktivní 3D grafiku a testové otázky k ověření pochopení probírané problematiky.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Sojka Petr doc. RNDr. Ph.D.**
Spojení	+420 549 494 377 sojka@mail.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Plch, Roman - Šarmanová, Petra - Sojka, Petr. Integrální počet funkcí více proměnných. Elportál, Brno : Masarykova univerzita, 2009, září, 160 s. ISSN 1802-128X.	D – článek ve sborníku (RIV 2009)	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/35/2009**

Název výsledku

Languages of Mathematics Random Walking in the Mathematics of Languages

Abstrakt

An essay about mathematics being a sublanguage of other natural languages: how it may be represented, stored, searched and handled in several projects of (European) Digital Mathematics Libraries as DML-CZ or EuDML. A framework for solving problem of computing of similar papers in a digital library is proposed, allowing several types of similarity type definitions: plagiarity counting on common word n-grams, topicality counting on common topics, or conarrativity counting on the same narrative. The vector of the most similar documents for a given similarity type is suggested to be computed using the algorithm by Page for web page ranking, often explained as `random walking"""""""""".

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Použití web page ranku pro výpočet vektoru popisujícího podobnost (plagiarita, topikalita a konaravitita) dokumentů.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Na základě vektoru podobnosti lze snadněji vyhledávat v databázi matematických článků (DML-CZ, EuDML, ...).

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Sojka Petr RNDr. Ph.D.**
Spojení	+420 549 494 377 sojka@mail.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Sojka, Petr. Languages of Mathematics Random Walking in the Mathematics of Languages. In RASLAN 2009 Proceedings. prvni. Brno : Masaryk University, 2009. od s. 127-133, 7 s. ISBN 978-80-210-5048-8.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/36/2009**

Název výsledku

Measuring Coverage of a Valency Lexicon using Full Syntactic Analysis

Abstrakt

Recent development showed that valency information provides a great benefit in many areas of natural language processing. Building valency lexicons is however a complex and time-consuming task from both theoretical and practical points of view, since designing of he lexicon plays a crucial role in its future usability as well as its careful and considerated preparation. As for any manually created resource, it is complicated to evaluate its quality. In this paper we consider the usage of the syntactic parser synt for estimating the coverage of the Verbalex verb valency lexicon for Czech. For this task we extended the phrase extraction functionality of the parser, which we describe briefly. Finally we discuss our results and further development.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Měření pokrytí valenčního slovníku (VerbaLex) pomocí syntaktické analýzy v korpusu.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Efektivnější výběr nových valenčních rámců pro přidání do valenčního slovníku.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Horák Aleš RNDr. Ph.D.**
Spojení	+420 549 494 377 hales@mail.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Jakubíček, Miloš - Kovář, Vojtěch - Horák, Aleš. Measuring Coverage of a Valency Lexicon using Full Syntactic Analysis. In RASLAN 2009 : Recent Advances in Slavonic Natural Language Processing. 1. vyd. Brno : Masaryk University, 2009. od s. 75-79, 5 s. ISBN 978-80-210-5048-8.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/37/2009**

Název výsledku

Mining Phrases from Syntactic Analysis

Abstrakt

In this paper we describe the exploitation of the syntactic parser synt to obtain information about syntactic structures (such as noun or verb phrases) of common sentences in Czech. The parser has been extended in such a way that enables its highly ambiguous output to be used for mining those phrases unambiguously and offers several ways how to identify them. Finally, an application for shallow valency extraction and punctuation correction is presented.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Využití výstupu syntaktického analyzátoru synt na extrakci syntaktických struktur (jmenné a slovesné fráze).

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Postup extrakce frází umožňuje povrchovou extrakci slovesných valencí a automatickou opravu interpunkce.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Horák Aleš RNDr. Ph.D.**
Spojení	+420 549 494 377 hales@mail.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Jakubíček, Miloš - Horák, Aleš - Kovář, Vojtěch. Mining Phrases from Syntactic Analysis. In Text, Speech, Dialogue 2009. 1. vyd. Berlin Heidelberg : Springer Verlag, 2009. od s. 124-130, 7 s. ISBN 978-3-642-04207-2.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/38/2009**

Název výsledku

Multilingual Features of Complex Valency Frames

Abstrakt

In this paper we deal with the verb valency lexicon of Czech verbs named VerbaLex, which contains complex valency verb frames (CVFs) including both surface and deep valencies. The most notable features of CVFs include two-level semantic labels with linkage to the Princeton and EuroWordNet Top Ontology hierarchy and the corresponding surface verb frame patterns capturing the morphological cases that are typical of the highly inflected languages like Czech. We discuss the assumption that CVFs are suitable for a description of the predicate-argument structure not only of Czech verbs but also verbs in other languages, particularly Bulgarian, Romanian and English. We come to the conclusion that this hypothesis can be verified reliably enough exploiting the principle of translatability and also indirectly using semantic classes of (Czech) verbs.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Využití slovesného lexikonu VerbaLex (popis struktury predikát-argument) nejen u českých sloves, ale i v jazycích jako bulharština, rumunština a angličtina.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Díky popsaným vlastnostem bude možné popisovat strukturu predikát-argument u sloves i ve výše zmíněných jazicích.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Horák Aleš RNDr. Ph.D.**
Spojení	+420 549 494 377 hales@mail.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Pala, Karel - Horák, Aleš. Multilingual Features of Complex Valency Frames. In Recent Advances in Intelligent Information Systems. Warsaw, Poland : Academic Publishing House EXIT, 2009. od s. 41-49, 9 s. ISBN 978-83-60434-59-8.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/39/2009**

Název výsledku

Problems of Machine Translation Evaluation

Abstrakt

In this article we deal with general aspects of machine translation evaluation. We describe several commonly used methods of the evaluation and discuss their problems and shortcomings. Then we outline a few thoughts and ideas which try to solve mentioned problems and stand behind a design of a new method of machine translation evaluation.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Popsaná nová metoda automatického hodnocení kvality strojového překladu využívá korpusu a slovníků a je nezávislá na ručně připravených referenčních překladech.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Popsaná metoda umožňuje hodnotit kvalitu libovolného překladu, nikli pouze omezené, předem připravené, množiny referenčních překladů.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Baisa Vít Mgr.**
Spojení	+420 777 592 171 xbaisa@fi.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Problems of Machine Translation Evaluation. In Proceedings of Recent Advances in Slavonic Natural Language Processing 2009. Brno : Masaryk University, 2009. 6 s. ISBN 978-80-210-5048-8.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/40/2009**

Název výsledku

Relations between Formal and Derivational Morphology in Czech

Abstrakt

The aim of the paper is to present results of the automatic analysis of some regular derivational types in Czech. In section 2 we briefly characterize the last obtained results in formal morphology of Czech (inflection) and its automatic processing as well as derivational morphology (word formation). In section 3 we show how the relations between word base and derived word can be described formally. In section 4 we explain how changes on the formal level correspond to semantic relations between related words (motivation relations). In section 5 we deal with the examples showing how the particular rules can be formulated and used in the new version of the software tool Deriv (derivational interface) that allows us to test derivational rules on the large machine dictionary of Czech stems. In section 6 we present particular results: rules which describe some selected derivational types. Finally, in section 7 we summarize the results obtained so far in the course of testing the software tool Deriv and indicate the further possibilities of its use.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- AI, 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Formální popis vztahů mezi slovník základem a výsledným slovem v procesu slovotvorby. Nová verze softwarového nástroje Deriv.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Automatická analýza pravidelných typů derivací v češtině zjednoduší práci při ručním návrhu derivačních vzorů.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Pala Karel doc. PhDr. CSc.**
Spojení	+420 549 495 616 pala@fi.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Osolsobě, Klára - Pala, Karel - Šmerk, Pavel - Hlaváčková, Dana. Relations between Formal and Derivational Morphology in Czech. In Czech in Formal Grammar. Mnichov : Lincom, 2009. od s. 79-87, 9 s. ISBN 978-3-89586-282-3.	D – článek ve sborníku (RIV 2009)	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/41/2009**

Název výsledku

Scaling to Billion-plus Word Corpora

Abstrakt

Most phenomena in natural languages are distributed in accordance with Zipf"s law, so many words, phrases and other items occur rarely and we need very large corpora to provide evidence about them. Previous work shows that it is possible to create very large (multi-billion word) corpora from the web. The usability of such corpora is often limited by duplicate contents and a lack of efficient query tools. This paper describes BiWeC, a Big Web Corpus of English texts currently comprising 5.5b words fully processed, and with a target size of 20b. We present a method for detecting near-duplicate text documents in multi-billion-word text collections and describe how one corpus query tool, the Sketch Engine, has been re-engineered to efficiently encode, process and query such corpora on low-cost hardware.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Tvorba miliardových korpusů, jejich efektivní tvorba a správa, odstraňování duplicit a provoz na levné výpočetní technice.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Použití velkých korpusových dat umožňuje postihnout i řídké jazykové jevy. Velké korpusy i pro tyto jevy poskytují evidenci, kterou běžně velké korpusy nemohou ze Zipfova zákona nikdy poskytnout.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Rychlý Pavel Mgr. Ph.D.**
Spojení	+420 549 496 399 pary@mail.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Pomikálek, Jan - Rychlý, Pavel - Kilgarriff, Adam. Scaling to Billion-plus Word Corpora. Advances in Computational Linguistics, Mexiko : Instituto Politécnico Nacional, 41, zima 2009, od s. 3-13, 14 s. ISSN 1870-4069. 2009.	J – článek v odborném periodiku (časopise) (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/42/2009**

Název výsledku

SCO 2009, Sharable Content Objects, 6th conference about electronic support of learning, Brno, Czech Republic

Abstrakt

Kniha obsahuje 37 recenzovaných příspěvků konference, 2 zvané přednášky, autorský, jmenný a tematický rejstřík. Kromě toho sborník obsahuje 20 anotací kurzů, studijních opor a technických řešení prezentovaných v paralelní sekci. Kniha je určena všem zájemcům o aktuální přístupy a výzkum v oblasti elektronické podpory výuky e-learningu.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Novinky v oblasti e-learningu jsou obsaženy v samotné knize, výběr: A Person Centered Approach to Including Students in Blended Learning, Stereometria a virtuálna realita, Podpora výuky matematické analýzy interaktivní 3D grafikou v PDF dokumentech, Interactive Support and Animation Tools in Mathematics Education, ...

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Podpora výuky mnoha popsanými prostředky se zaměřením na e-learning.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Sojka Petr doc. RNDr. Ph.D.**
Spojení	+420 549 494 377 sojka@mail.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Sojka, Petr - Rambousek, Jiří. SCO 2009, Sharable Content Objects, 6. ročník konference o elektronické podpoře výuky, Brno, Česká republika. Edited by Sojka P., Rambousek J. prvni. Brno : Masaryk University, 2009. 296 s. SCO Proceedings. ISBN 978-80-210-4878-2.	B – odborná kniha (RIV 2009)	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/43/2009**

Název výsledku

SCO 2009, Sharable Content Objects, 6th conference about electronic support of learning, Brno, Czech Republic

Abstrakt

Na 6. ročníku konference o elektronické podpoře výuky bylo prezentováno 37 recenzovaných příspěvků konference, 2 zvané přednášky a 20 anotací kurzů a studijních opor prezentovaných v paralelní sekci. Konference byla určena všem zájemcům o aktuální přístupy a výzkum v oblasti elektronické podpory výuky e-learningu.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Novinky v oblasti e-learningu jsou obsaženy v samotné knize, výběr: A Person Centered Approach to Including Students in Blended Learning, Stereometria a virtuálna realita, Podpora výuky matematické analýzy interaktivní 3D grafikou v PDF dokumentech, Interactive Support and Animation Tools in Mathematics Education, ...

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Podpora výuky mnoha popsanými prostředky se zaměřením na e-learning.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Sojka Petr doc. RNDr. Ph.D.**
Spojení	+420 549 494 377 sojka@mail.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/44/2009**

Název výsledku

Shallow Ontology Based on VerbaLex

Abstrakt

Ontologies have proven to be a useful resource in natural language processing. In this paper, we introduce basic ideas of a shallow ontology named Sholva. This ontology is based on VerbaLex, a database of verb valencies, where each valency pointer also contains a pointer into EuroWordnet. We focused our effort on building ontology which would help us in solving real problems in syntactic analysis, word sense disambiguation and machine translation.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Využití EuroWordnetu a VerbaLexu při tvorbě ontologie Sholva.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Ontologie Sholva pomáhá řešit reálné problémy v syntaktické analýze, určování významu slov a strojovém překladu.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Grác Marek Mgr.**
Spojení	+420 549 494 377 xgrac@fi.muni.cz
Organizace	00216224 Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Grác, Marek. Shallow Ontology Based on VerbaLex. In NLP, Corpus Linguistics Corpus Based Grammar Research. Bratislava, SR : Slovenská akadémia vied, Jazykovedný ústav Ľ. Štúra, 2009. 400 s. ISBN 978-80-7399-875-2.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/45/2009**

Název výsledku

Syntactic Analysis as Pattern Matching: The SET Parsing System

Abstrakt

In this paper, we show a new approach to syntactic analysis of free-word-order languages based on the idea of pattern matching linking rules. The system, named SET, is currently developed and tested with the Czech language as a representative of free-word-order languages with very rich morphological system.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Nový přístup k syntaktické analýze jazyků s bohatou syntaxí s využitím pattern matching (hledání vzorů).

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Jednoduchá správa gramatických pravidel, jednoduchý, snadno čitelný a adjustovatelný syntaktický analyzátor SET.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Kovář Vojtěch Mgr.**
Spojení	+420 549 494 377 vojcek@mail.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Kovář, Vojtěch - Horák, Aleš - Jakubíček, Miloš. Syntactic Analysis as Pattern Matching: The SET Parsing System. In Proceedings of 4th Language & Technology Conference. Poznań (Poland) : Wydawnictwo Poznańskie, 2009. od s. 100-104, 5 s. ISBN 978-83-7177-746-2.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/46/2009**

Název výsledku

Third Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2009

Abstrakt

The RASLAN workshop is an event dedicated to exchange of information between research teams working on projects of computer processing of Slavonic languages and related areas. RASLAN is focused on theoretical as well as technical aspects of the project work, presentations of verified methods are welcomed together with descriptions of development trends. The workshop also serves as a place for discussion about new ideas. The proceedings contain 17 contributions, written for researchers and advanced students interested in computational linguistic research of Slavonic languages (text corpora and tagging, syntactic parsing, sense disambiguation, semantic networks and ontologies, knowledge representation and applied systems and software).

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Inovační aspekty viz jednotlivé příspěvky ve sborníku.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Přínosy viz jednotlivé příspěvky ve sborníku.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Horák Aleš RNDr. Ph.D.**
Spojení	+420 549 494 377 hales@mail.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Sojka, Petr - Horák, Aleš. Third Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2009. Edited by Sojka P., Horák A. Brno : Masaryk University, 2009. 143 s. RASLAN Proceedings, Third (2009). ISBN 978-80-210-5048-8.	B – odborná kniha (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/47/2009**

Název výsledku

Trdlo, an Open Source Tool for Building Transducing Dictionary

Abstrakt

This paper describes the development of an open-source tool named Trdlo. Trdlo was developed as part of our effort to build a machine readable dictionary for Czech-Slovak language. Proposed methods describes in this paper attempt to extend existing dictionary with inferable translation pairs.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Vytváření automaticky odvoditelných překladových dvojic v česko-slovenském překladovém slovníku na základě morfologické analýzy a korpusových dat.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Trdlo umožňuje automatické rozšíření existujícího slovníku o odvoditelné překladové dvojice.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Grác Marek Mgr.**
Spojení	+420 549 494 377 xgrac@fi.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Grác, Marek. Trdlo, an Open Source Tool for Building Transducing Dictionary. In Text, Speech and Dialogue 2009, 12th International Conference TSD 2009. Berlin : Springer-Verlagen, 2009. od s. 64-69, 428 s. ISBN 3-642-04207-4.	D – článek ve sborníku (RIV 2009)	BUL

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/48/2009**

Název výsledku

Using Wordnets and Ontologies for Text-Meaning Assignment - Implementation Details of the KYOTO Project First Phase

Abstrakt

The vision of Semantic Web introduced ontologies as the main unifying tool for management of the knowledge and semantic structure of text documents. However, linking the real text documents with the ontologies (of various kinds and various degree of complexity) is still a matter of current research in knowledge representation projects. In this paper, we are presenting the work results of the KYOTO project database implementation. The goal of the project is to provide a complex system for automatic processing of documents in order to extract known facts, link them with shared ontology and use this knowledge for Question Answering about the document topic. We give details about the design and implementation of the KYOTO database, which interlinks national WordNet semantic networks with the general SUMO ontology to offer the basis of the future shared ontology.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Propojení sémantické sítě WordNetu s obecnou ontologií SUMO a vytvoření základu pro budoucí sdílené ontologie.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Cílem je poskytnout komplexní systém pro automatické zpracování dokumentů a extrakci známých faktů pomocí (mimojiné) těchto ontologií.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Horák Aleš RNDr. Ph.D.**
Spojení	+420 549 494 377 hales@mail.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Horák, Aleš - Rambousek, Adam. Using Wordnets and Ontologies for Text-Meaning Assignment - Implementation Details of the KYOTO Project First Phase. In Proceedings of the 4th International Conference on Software and Data Technologies, Volume 2. Portugalsko : INSTICC, 2009. od s. 303-307, 5 s. ISBN 978-989-674-010-8.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/49/2009**

Název výsledku

Yet Another Formalism for Morphological Paradigm

Abstrakt

Morphology is one of the few areas in the natural language processing where computers are good enough. Different approaches lead to different problems. For Slavonic languages rules and statistical methods are commonly used. Rule based methods are more precise but tend to fail when parsing unknown words. Hybrid technologies with statistical methods helps to solve this problem. It is also possible to solve this problem by extending existing rule-based resources. These resources can be used also for other linguistic research. This paper presents new formalism which is closer to human understanding of natural language morphology and its application in extending morphological dictionary.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Nový formalismus pro morfologii, který je blíže lidskému chápání morfologie přirozeného jazyka.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Řešení problémů spojených s dvěma odlišnými principy morfologické analýzy: pravidlové a statistické, použitím hybridního přístupu.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Grác Marek Mgr.**
Spojení	+420 549 494 377 xgrac@fi.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Grác, Marek. Yet Another Formalism for Morphological Paradigm. In Proceedings of RASLAN 2009. Brno : Masaryk University, 2009. 134 s. ISBN 978-80-210-5048-8.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/50/2009**

Název výsledku

The Influence of Text Pre-processing on Plagiarism Detection

Abstrakt

This paper explores the influence of text pre-processing techniques on plagiarism detection. We examine stop-word removal, lemmatization, number replacement, synonymy recognition, word generalization, and various combinations of these techniques. We also look into the influence of punctuation and word-order within N-grams. All these techniques are evaluated according to their impact on F1-measure and speed of execution. Our experiments were performed on a Czech corpus of plagiarized documents about politics. At the end of this paper, we propose what we consider to be the best combination of text pre-processing for plagiarism detection.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- BD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Využití WordNet tezauru v oblasti odhalování plagiátů. V psaném textu jsou rozpoznávána synonyma a zobecňovány významy slov, tzv. vyhledávání vhodných hyperonym.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Zvýšení výpočetní rychlosti při zpracovávání generalizovaného, méně objemného textu. Aplikace těchto technik neovlivňuje přesnost detekce plagiátů, v některých případech dochází dokonce ke zlepšení výsledků.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Češka Zdeněk Ing. PhD.**
Spojení	377 632 452 377 632 402 zceska@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň www.textmining.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Ceska, Z., Fox, C.: „The Influence of Text Pre-processing on Plagiarism Detection“. In: Proceedings of the 13th International Conference on Recent Advances in Natural Language Processing, pp. 55-59, Borovets, Bulgaria, September 2009. ISSN 1313-8502.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/51/2009**

Název výsledku

Využití techniky náhodného indexování v oblasti detekce plagiátů

Abstrakt

Rostoucí snaha plagiovat cizí práce, především v oblasti školství, zapříčinila vývoj nových a lepších metod, které by těmto intrikám čelily. Tento článek rozvíjí myšlenku aplikace Latentní sémantické analýzy (LSA) v oblasti detekce plagiátů a navrhuje nová vylepšení. Hlavním diskutovaným předmětem je aplikace kompresní techniky, tzv. náhodného indexování, která transformuje data do alternativního zmenšeného prostoru. Kromě toho se článek zabývá normalizací podobností mezi dokumenty a přináší novou asymetrickou normalizační formuli. Experimenty byly provedeny na manuálně vytvořeném korpusu českých plagiátů, který obsahuje 1500 dokumentů o politice. Dosažené výsledky indikují, že kompresní technika dokáže významně snížit časové požadavky pro LSA. Aplikací nové normalizační formule lze navíc dosáhnout i vyšší přesnosti detekce plagiátů při současně nižších časových požadavcích.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- BD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Aplikace kompresní techniky náhodného indexování v oblasti odhalování plagiátů.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Podstatným způsobem komprimuje matici příznaků, která reprezentuje model dokumentů. Aplikací tohoto postupu se významně snižují časové požadavky na výpočet latentní sémantické analýzy, při současném zachování dobrých výsledků v přesnosti odhalování plagiátů.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Česka Zdeněk Ing. PhD.**
Spojení	377 632 452 377 632 402 zceska@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň www.textmining.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Češka, Z.: „Využití techniky náhodného indexování v oblasti detekce plagiátů“. In: Proceedings of the ITAT 2009, Information Technologies - Applications and Theory, pp. 23-26, Kralova Studna, Slovakia, September 2009. ISBN 978-80-970179-1-0.	D – článek ve sborníku (RIV 2009)	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/52/2009**

Název výsledku

Porovnání technik předzpracování textu pro detekci plagiátů

Abstrakt

Článek se zabývá technikami předzpracování textu a jejich vlivem na detekci plagiátů v psaném textu. V našich experimentech zkoumáme stop slova, lemmatizaci, nahrazování synonym a jejich vzájemné kombinace. Dále navrhujeme pokročilou normalizaci slov s využitím hyperonym z WordNet tezauru. Testy jsme provedli na českém korpusu plagiátů čítajícím 950 dokumentů o politice, vytvořeném z ČTK korpusu. Pro experimenty používáme metodu postavenou na RFM, prostém srovnání N gramů s Jaccard-Tanimoto koeficientem a metodu pracující na principu singulární dekompozice vztahů frází.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- BD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Vliv rozličných technik pro předzpracování textu na přesnost odhalování plagiátů.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Zvýšení výpočetní rychlosti metody pro odhalování plagiátů. Výsledků je dosaženo zpracováváním menšího objemu sémanticky významného textu, který je filtrován technikami předzpracování.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Češka Zdeněk Ing. PhD.**
Spojení	377 632 452 377 632 402 zceska@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň www.textmining.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Češka, Z.: „Porovnání technik předzpracování textu pro detekci plagiátů“. In: Proceedings of the 8th Annual Conference ZNALOSTI 2009, Brno, Czech Republic, pp. 293-296, February 2009. ISBN 978-80-227-3015-0.	D – článek ve sborníku (RIV 2009)	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/53/2009**

Název výsledku

Návrh a implementace pokročilé metody pro odhalování plagiátů s využitím LSA

Abstrakt

Implementace pokročilé metody pro odhalování plagiátů, která využívá principů latentní sémantické analýzy. Součástí této implementace jsou rovněž techniky pro předzpracování textu a kompresi příznaků. Implementace je formou DLL knihovny v jazyce C# (.NET Framework 3.5). Navržená metoda byla implementována a ověřena na kolekci 1500 českých plagiovaných textových dokumentů.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- BD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Aplikace latentní sémantické analýzy v oblasti odhalování plagiátů a zapojení kompresních technik.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Inovativní metoda pro odhalování plagiátů, která dosahuje lepších výsledků v porovnání s ostatními.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Češka Zdeněk Ing. PhD.**
Spojení	377 632 452 377 632 402 zceska@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň www.textmining.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Dokumentace k software SVDPlag v.1.0 je dostupná na http://www.kiv.zcu.cz/vyzkum/software/	R – software (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/54/2009**

Název výsledku

EuroSearch - vyhledávání v multilinguálním prostředí

Abstrakt

S postupující integrací jednotlivých států a rozšiřováním Evropské unie se dostává respektování vícejazyčnosti do popředí zájmu. Typickým příkladem aplikace multinguálního systému může být prohledávání webových stránek, vědeckých článků, zákonů, předpisů a podobně. Lze také rozšířit stávající vyhledávací systémy tak, aby lépe umožňovaly vyhledávání ve vícejazykovém prostředí. Vytvořený systém nachází uplatnění v rozsáhlejších digitálních knihovnách, kde se vyskytují dokumenty v různých jazycích, případně ve státní správě, která bude stále častěji přicházet do styku s cizojazyčnými dokumenty. Za předpokladu, že uživatel zná několik jazyků, je vhodné umožnit jedním dotazem vyhledat více relevantních dokumentů.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- BD, 2.- JD, 3.- AI, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Byl vytvořen softwarový balík umožňující prohledávání webových stránek, vědeckých článků, zákonů, předpisů a podobně. Lze jím rozšířit stávající vyhledávací systémy tak, aby lépe umožňovaly vyhledávání ve vícejazykovém prostředí. Vytvořený systém nachází uplatnění v rozsáhlejších digitálních knihovnách, kde se vyskytují dokumenty v různých jazycích, a ve státní správě.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Zpracovaný software dovoluje vyhledávání dokumentů ve vícejazyčné kolekci za použití EWN. Pracuje s anglickým a s českým textem, rozšíření na další jazyky je možné úpravou lematizačních tabulek a slovníku EuroWordnetu.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Toman Michal Ing.**
Spojení	377632479 777575983 377632402 mtoman@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň www.textmining.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Software, uložen na http://www.kiv.zcu.cz/vyzkum /software/	R – software (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/55/2009**

Název výsledku

Automatic Plagiarism Detection Based on Latent Semantic Analysis

Abstrakt

Plagiarism is a widely spread problem that is the main focus of interest these days. The main objective of this PhD thesis is the application of Latent Semantic Analysis (LSA) framework in the field of written-text plagiarism detection. This particular field faces various issues that are discussed thoroughly. In order to infer the latent semantics from the given text, Singular Value Decomposition (SVD) is employed for the purpose of large statistical computations. That is why the proposed method is called SVDPlag. To overcome issues connected with a large amount of extracted N-grams from the text, a feature selection and subsequently a random indexing techniques are applied. Moreover, this thesis deals with the influence of text pre-processing on the accuracy of plagiarism detection. Simultaneously, the aspects of multilingual environment are explored. Various approaches in common use are discussed and compared with the new proposed method. A Czech corpus of 1,500 text documents about politics – created manually by students – was employed for the experiments. The results indicate that SVDPlag method significantly improves the accuracy of plagiarism detection and outperforms the other methods.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- BD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Aplikace latentní sémantické analýzy v oblasti odhalování plagiátů, vliv předzpracování textu a možnosti vícejazyčného zpracování.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Nová inovativní metoda SVDPlag, založená na latentní sémantické analýze, která překonává ostatní přístupy v přesnosti při odhalování plagiátů. Kromě toho jsou rozebírány techniky předzpracování textu, vícejazyčné prostředí a jejich celkový dopad na schopnost správně identifikovat plagiovaný text. Jedná se o souhrnnou publikaci, jež prezentuje výsledky metody SVDPlag a srovnává je s ostatními existujícími metodami.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Češka Zdeněk Ing. PhD.**
Spojení	377 632 452 377 632 402 zceska@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň www.textmining.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Češka, Z.: Automatic Plagiarism Detection Based on Latent Semantic Analysis. PhD Thesis, KIV ZČU v Plzni, 2009		ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/56/2009**

Název výsledku

Detection and Annotation of Graphical Objects in Raster Images within the GATE Project.

Abstrakt

Článek se zabývá problematikou detekce a anotace grafických objektů v rastrových obrázcích, zejména možnostmi nalezení hranice objetů a anotací v rámci projektu GATE (Graphics Accessible for Everyone. Zahrnuje rovněž ilustrativní případ navrženého postupu.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Článek navrnuje inovativní meody detekce grafických objektů v systému GATE.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Přínos výsleku spočívá v možnosti urychlení anotačního procesu grafických objektů.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Kopeček Ivan Doc. RNDr. CSc.**
Spojení	549493861 731268249 549491820 kopeček@fi.muni.cz
Organizace	00216624 Fakulta informatiky MU Botanická 68a 60200 Brno http://www.fi.muni.cz/

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Kopeček, Ivan - Ošlejšek, Radek - Plhák, Jaromír - Tiršel, Fedor. Detection and Annotation of Graphical Objects in Raster Images within the GATE Project. In Proceedings of the 2009 International Conference on Internet Computing ICOMP 2009. USA : CSREA Press, 2009. od s. 285-290, 6 s. ISBN 1-60132-110-4.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/57/2009**

Název výsledku

Accessibility of Graphics and E-learning.

Abstrakt

Článek se zabývá problematikou přístupnosti grafiky pro nevidomé a možnostmi využití nově navrhovaných přístupů a metod pro e-learning.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Inovativním aspektem je použií metod pro zpřístupnění grafiky pro nevidomé v e-lerningu.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Popsané medrody dovolují nevidomým studentům pracovat s grafickými objekty jako jsou fotografie, obrázky, schemata apod.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Kopeček Ivan Doc. RNDr. CSc.**
Spojení	549493861 731268249 549491820 kopecek@fi.muni.cz
Organizace	00216224 Fakulta informatiky MU Botanická 68a 60200 Brno http://www.fi.muni.cz/

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Kopeček, Ivan - Ošlejšek, Radek. Accessibility of Graphics and E-learning. In Proceedings of the Second International Conference on ICT & Accessibility. Hammamet : Art Print, 2009. od s. 157-165, 9 s. ISBN 978-9973-37-516-2.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/58/2009**

Název výsledku

Generating Dialogues from the Description of Structured Data

Abstrakt

Generating dialogues from the description of structured data is a part of the WebGen system for generating web-based presentations by means of dialogue. Although the system allows the user to create the most common presentations types, such as personal presentations or blogs, the user may need to add a new presentation type. To add a new presentation type the user must specify the content and structure of the presentation descriptor using the XML Schema, the dialogue interface used to collect requested information and the layout of the resulting pages. This paper discusses limitations on the XML Schema structure allowing us to generate dialogue interfaces automatically from the presentation descriptor and basic principles and algorithms used during the transformation. Illustrative examples of the data, corresponding XML Schema, XSL Transformation and resulting VoiceXML document is presented as well.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Umožňuje přidávat do systému dialogová rozhraní pouze na základě popisu dat, která mají být získána.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Možnost automatického doplnění dialogového rozhranní do systému bez znalosti používaných standardů pouze na základě dodaného popisu dat, která mají být získána.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Bártek Luděk Mgr. Ph.D.**
Spojení	+420549493215 bar@fi.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Generating Dialogues from the Description of Structured Data. In HCI and Usability for e-Inclusion. Heidelberg : Springer-Verlag, 2009. od s. 227-235, 9 s. ISBN 978-3-642-10307-0.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/59/2009**

Název výsledku

Web Mining Methods for the Detection of Authoritative Sources: Theory and Practice

Abstrakt

The development of information society in recent decades has enabled collecting, filtering and storing huge amounts of data. These data must be further processed to gain valuable information and knowledge. The scientific field dealing with extracting information and knowledge from data has evolved rapidly to cope with the extent and growth of information sources the number of which has geometrically increased with the appearance of the World Wide Web. All traditional approaches in information retrieval, knowledge acquisition, and data mining must be adapted for the dynamic, heterogeneous, and unstructured data on the Web. Web mining has come into being as a fully-fledged research discipline. This book presents state-of-the-art knowledge of Web mining from the perspective of looking for authoritative sources. Besides introduction to the theoretical concepts of Web crawling, ranking algorithms, and social networks, results of practical experiments are shown as well. In particular, a brand new algorithm for bibliographic networks is introduced. This publication will be especially useful to professionals, researchers, and students in the field of data mining and information retrieval.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- BD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Byly shrnuty soudobé poznatky o principech a metodách web miningu a data miningu a byla vyvinuta zcela nová metodologie pro hodnocení významnosti uzlů v orientovaných grafech, zejména v citačních sítích. Informace z citačních sítí jsou nově obohaceny o další informace z grafů spolupráce.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Techniky využité v rámci aktivity představují v českém prostředí dosud nedoceněné případy strojového zpracování webových dat za účelem hodnocení kvality vědeckého výzkumu.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Fiala Dalibor Ing. PhD.**
Spojení	377 632 429 723 263 439 377 632 402 dalfia@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň www.textmining.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Fiala, D.: Web Mining Methods for the Detection of Authoritative Sources: Theory and Practice. VDM Verlag, Saarbrücken, Germany, 2009. ISBN 978-3639173376	B – odborná kniha (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/60/2009**

Název výsledku

Extended Formal Model for Ranking of Authoritative Resources

Abstrakt

The aim of this paper is to inform how authority of researchers can be detected from information hidden in citation networks. The citation networks are networks of relationships between citing and cited authors, publications and other similar resources (e.g. hyperlinked structures of the Web). We want to derive a rating of individual participants modeled as nodes of this network. When the publication-citation graph is enlarged and weights of edges expressing the time of citation are added, we have the chance to order author-author citations in time. Consequently, there is possible to decrease the weights of citations between authors when citations are cycling.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- BD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Je navržen princip vyhodnocování autoritativnosti vědeckých pracovníků modifikací metody PageRanku, zohledňující jak spoluautorství, tak i časové hledisko citací. Snižováním vah hran v citačním grafu umožní eliminovat vliv případných citačních loby.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Citační cykly zkreslují výslednou autoritativnost. Jejich zohledněním získáme ohodnocení významnosti lépe odpovídající realitě.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Ježek Karel Doc. Ing. CSc.**
Spojení	377 632 475 377 632 402 jezek_ka@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 2732 8 30614 Plzeň www.textmining.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Jezek K.: Extended Formal Model for Ranking of Authoritative Resources. Proccedings XXII Mezdunarodnoj Naucnoj Konferencii MMTT-22, (Tom 7), pp.193-195, Pskov 2009. ISBN 978-5-91116- 087-2	D – článek ve sborníku (RIV 2009)	ANG

4.1.3. PLNĚNÍ DÍLČÍCH CÍLŮ

4.1.3.1. ZPRÁVA O DOSAŽENÍ DÍLČÍHO CÍLE

Číslo dílčího cíle	3
Název dílčího cíle	Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce.
Plánované datum dosažení dílčího cíle	31.12.2009

INDIKÁTORY DOSAŽENÍ VÝSTUPU - SKUTEČNĚ DOSAŽENÉ

Výsledkem snah řešitelů je vytvoření celé skupiny unikátních metod umožňujících efektivní úpravy a využívání webových stránek s moderní strukturou, komunikaci s webovými stránkami poměrně rozsáhlou podmnožinou přirozeného jazyka, dokončení tvorby potřebných datových kolekcí a tvorby různých datových korpusů. V rámci této aktivity dále bylo přistoupeno k ověřování některých prototypových řešení. Lze konstatovat, že až na některé drobné detaily byla tato fáze řešení projektu naplněna, ovšem některé činnosti z tohoto úkolu budou i nadále doplňovány. Především půjde o metody automatického rozpoznávání dialogových aktů na základě dialogové historie a dokončení tématu s názvem „Extrakce informací z textů e-mailů typu Call for papers (žádost o zaslání příspěvků na konferenci)“, do kterého budou doplněny některé rozvinutější dialogy, jejichž programová implementace vyžaduje náročná programová řešení.

PROSTŘEDKY OVĚŘENÍ VÝSTUPU - SKUTEČNĚ DOSAŽENÉ

Obdržené výsledky budeme v dalším období rozsáhle testovat a předpokládáme ověření funkčnosti dialogového systému na dialozích získaných jednak doslovným přepisem zaznamenaných dialogů a prezentací a poté též vedením přímých dialogů třemi kategoriemi uživatelů – řešiteli a spoluřešiteli projektu, kolegy a studenty katedry nezúčastněnými na řešení projektu a konečně náhodně zastavenými osobami, které dosud s takovým systémem neměly možnost kontaktu. Výsledky dialogů budou kompletně zaznamenávány a poté detailně vyhodnocovány na základě vyhodnocování zejména dialogů třetí skupiny osob (vedených osobami neznalými) bude systém následně upravován a jeho funkce korigovány. K testování prototypových systémů budou i nadále využívány upravené a doplněné korpusy, pro vyhledávání a sumarizaci dokumentů a klasifikaci dokumentů budou využívány nově implementované metody.

4.1.4. REDAKČNĚ UPRAVENÁ ZPRÁVA

Při řešení projektu v roce 2009 se řešitelé zaměřili na splnění dílčího úkolu označeného číslem 3 – návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. Cílem bylo vytvoření celé skupiny unikátních metod umožňujících efektivní úpravy a využívání webových stránek s moderní strukturou, komunikaci s webovými stránkami poměrně rozsáhlou podmnožinou přirozeného jazyka, dokončení tvorby potřebných datových kolekcí a tvorby různých datových korpusů. Lze konstatovat, že až na některé drobné detaily byla třetí fáze řešení projektu naplněna, ovšem některé činnosti z tohoto úkolu budou i nadále doplňovány. Především půjde o metody automatického rozpoznávání dialogových aktů na základě dialogové historie a dokončení tématu s názvem „Extrakce informací z textů e-mailů typu Call for papers (žádost o zaslání příspěvků na konferenci)“, do kterého budou doplněny některé rozvinutější dialogy, jejichž programová implementace vyžaduje náročná programová řešení. V příštím období budou výsledky obsáhle testovány a předpokládáme ověření funkčnosti dialogového systému na dialozích získaných jednak doslovným přepisem zaznamenaných dialogů a prezentací a též vedením přímých dialogů třemi kategoriemi uživatelů – řešiteli a spoluřešiteli projektu, kolegy a studenty katedry nezúčastněnými na řešení projektu a konečně náhodně zastavenými osobami, které dosud s takovým systémem neměly možnost kontaktu. Výsledky dialogů budou kompletně zaznamenávány a poté detailně vyhodnocovány na základě vyhodnocování zejména dialogů třetí skupiny osob (vedených osobami neznalými) bude systém následně upravován a jeho funkce korigovány. A konečně byly během roku 2009 vytvořeny další verze systému, s nimiž byly testovány rozsáhlejší soubory dokumentů a podařilo se jimi ověřit základní myšlenky podcílů projektu.

4.1.5. PLNĚNÍ PODMÍNEK PROGRAMU

Plnění specifických podmínek programu - se pro projekty NPV II nezpracovává. Pro projekty NPVII specifické podmínky ve vyhlášení programu nebyly formulovány.

4.1.6. PLNĚNÍ SMLOUVY O SPOLUPRÁCI

Na základě vymezených základních práv (viz uzavřená smlouva upravující vztahy mezi příjemcem a spolupříjemcem) příjemce poskytnul spolupříjemci finanční dotaci přímým převodem na stanovený účet Masarykovy univerzity, náklady na projekt byly vedeny v oddělené evidenci obou spolupracujících subjektů.
Uzavřená smlouva o spolupráci je plněna beze zbytku, plánované finanční prostředky byly vyčerpány - viz odstavec 2.3.2.

4.2. DALŠÍ PŘÍLOHY - rok 2009

4.2.1. Odborné a věcné přílohy zprávy - seznam

	Pořadí	Soubor
	1	**Seznam publikovaných prací - ZČU v Plzni** Soubor obsahuje výčet všech publikací, které vznikly v rámci řešení projektu 2C06009 v průběhu roku 2009 Seznam_publikaci_2009.doc *(40 kB)*
	2	**Seznam publikací - MU Brno** Seznam 28 publikací MU seznam-publikaci.doc *(163 kB)*

4.2.2. Ostatní (např. možné využití výsledků) - seznam

	Pořadí	Soubor
	1	**Vybrané publikace - Plzeň** Zazipovaný soubior 15 nejvýznamnějších publikací z roku 2009 Publikace_Plzen.zip *(2721 kB)*
	2	**Vybrané publikace - Brno** Vybrané významné publikace z Masarykovy univerzity v Brně Publikace_Brno.zip *(980 kB)*

4.2.3. Zápisy z projednání (oponentní řízení, atd.) - seznam

	Pořadí	Soubor
	1	**Oponentní řízení** Oponentní řízení nebylo v roce 2009 konáno. *(kB)*

4.2.4. Zápisy a dokumenty z jednání s administrátory programu poskytovatele - seznam

	Pořadí	Soubor
	1	**Pracovní jednání s pracovníky poskytovatele** Pracovní jednání se styčnými pracovníky poskytovatele se v roce 2009 neuskutečnila. Pouze byla písemně dohodnuta drobná změna v čerpání finančních prostředků pro řešitele z důvodu změny struktury řešitelských týmů a financování účastnických poplatků na konferencích. *(kB)*

4.2.5. Zápisy z jednání Rady projektu (Centra) - seznam

Příloha 4.2.5. Zápisy z jednání Rady projektu (Centra) - se pro tento program nezpracovává.

4.2.6. Návrh dodatku ke smlouvě na řešení projektu se zdůvodněním - seznam

Příloha 4.2.6. Návrh dodatku ke smlouvě na řešení projektu se zdůvodnění - se pro tento program nezpracovává.